Feature-Augmented Neural Networks for Patient Note De-identification

Ji Young Lee[1*], **Franck Dernoncourt**[1*], **Özlem Uzuner**[2], **Peter Szolovits**[1]
[1]MIT, [2]SUNY Albany
{jjylee,francky}@mit.edu, ouzuner@albany.edu, psz@mit.edu
[*] These authors contributed equally to this work.

Abstract

Patient notes contain a wealth of information of potentially great interest to medical investigators. However, to protect patients' privacy, Protected Health Information (PHI) must be removed from the patient notes before they can be legally released, a process known as patient note de-identification. The main objective for a de-identification system is to have the highest possible recall. Recently, the first neural-network-based de-identification system has been proposed, yielding state-of-the-art results. Unlike other systems, it does not rely on human-engineered features, which allows it to be quickly deployed, but does not leverage knowledge from human experts or from electronic health records (EHRs). In this work, we explore a method to incorporate human-engineered features as well as features derived from EHRs to a neural-network-based de-identification system. Our results show that the addition of features, especially the EHR-derived features, further improves the state-of-the-art in patient note de-identification, including for some of the most sensitive PHI types such as patient names. Since in a real-life setting patient notes typically come with EHRs, we recommend developers of de-identification systems to leverage the information EHRs contain.

1 Introduction and related work

Medical practitioners increasingly store patient data in Electronic Health Records (EHRs) (Hsiao et al., 2011), which represents a considerable opportunity for medical investigators to construct novel models and experiments to improve patient care. Some governments even subsidize the adoption of EHRs, such as the Centers for Medicare & Medicaid Services in the United States who have spent over $30 billion in EHR incentive payments to hospitals and medical providers (McCann, 2015).

A legal prerequisite for a patient note to be shared with a medical investigator is that it must be de-identified. The objective of the de-identification process is to remove all Protected Health Information (PHI). Not appropriately removing PHI may result in financial penalties (DesRoches et al., 2013; Wright et al., 2013). In the United States, the Health Insurance Portability and Accountability Act (HIPAA) (Office for Civil Rights, 2002) defines PHI types that must be removed, ranging from phone numbers to patient names. Failure to accurately de-identify a patient note would jeopardize the patient's privacy: the performance of a de-identification system is therefore critical.

A naive approach to de-identification is to manually identify PHI. However, this is costly (Douglass et al., 2005; Douglas et al., 2004) and unreliable (Neamatullah et al., 2008). Consequently, there has been much work developing automated de-identification systems. These systems are either based on rules or machine-learning models. Rule-based systems typically rely on patterns, expressed as regular expressions and gazetteers, defined and tuned by humans (Berman, 2003; Beckwith et al., 2006; Fielstein et al., 2004; Friedlin and McDonald, 2008; Gupta et al., 2004; Morrison et al., 2009; Neamatullah et al., 2008; Ruch et al., 2000; Sweeney, 1996; Thomas et al., 2002).

Machine-learning-based systems train a classifier to label each token as PHI or not PHI. Some systems are more fine-grained by detecting which PHI type a token belongs to. Different statistical methods have been explored for patient note de-identification, including decision trees (Szarvas et al., 2006), log-linear models, support vector machines (SVMs) (Guo et al., 2006; Uzuner et al., 2008; Hara, 2006), and conditional random field (CRFs) (Aberdeen et al., 2010). A thorough review of existing systems can be found in (Meystre et al., 2010; Stubbs et al., 2015).

A more recent system has introduced the use of artificial neural networks (ANNs) for de-identification (Dernoncourt et al., 2016), and obtained state-of-the-art results. The system does not use any manually-curated features. Instead, it solely relies on character and token embeddings. While this allows the system to be developed and deployed faster, it fails to give users the possibility to add features engineered by human experts. Additionally, in practical settings of de-identification, patient notes typically come from a hospital EHR database, which contains metadata such as which patient each note pertains to, and other information such as the names of all doctors who work at the hospital where the patient was treated. The features derived from EHR databases may be useful for boosting the performance of de-identification systems. In this work, we present a method to incorporate features to this ANN-based system, and show that it further improves the state-of-the-art.

2 Method

The first model based on ANNs for patient note de-identification was introduced in (Dernoncourt et al., 2016): we extend upon their model. They utilized both token and character embeddings to learn effective features from data by fine-tuning the parameters. The main components of the ANN model are Long Short Term Memories (LSTMs) (Hochreiter and Schmidhuber, 1997), which are a type of recurrent neural networks (RNNs).

The model is composed of three layers: a character-enhanced token embedding layer, a label prediction layer, and a label sequence optimization layer. The character-enhanced token embedding layer maps each token into a vector representation. The sequence of vector representations corresponding to a sequence of tokens are input to the label prediction layer, which outputs the sequence of vectors containing the probability of each label for each corresponding token. Lastly, the sequence optimization layer outputs the most likely sequence of predicted labels based on the sequence of probability vectors from the previous layer. All layers are learned jointly. For more details on the basic ANN model, see (Dernoncourt et al., 2016).

We augment this ANN model by adding features that are human-engineered or derived from EHR database, as presented in Table 1. The majority of human-engineered features are taken from (Filannino and Nenadic, 2015), a few more features come from (Yang and Garibaldi, 2015), and additional gazetteers are collected using online resources. All features are binary and computed for each token. The binary feature vector comprising all features for a given token is fed into a feedforward neural network, the output vector of which is concatenated to the corresponding token embeddings, at the output of the character-enhanced token embedding layer, as Figure 1 illustrates.

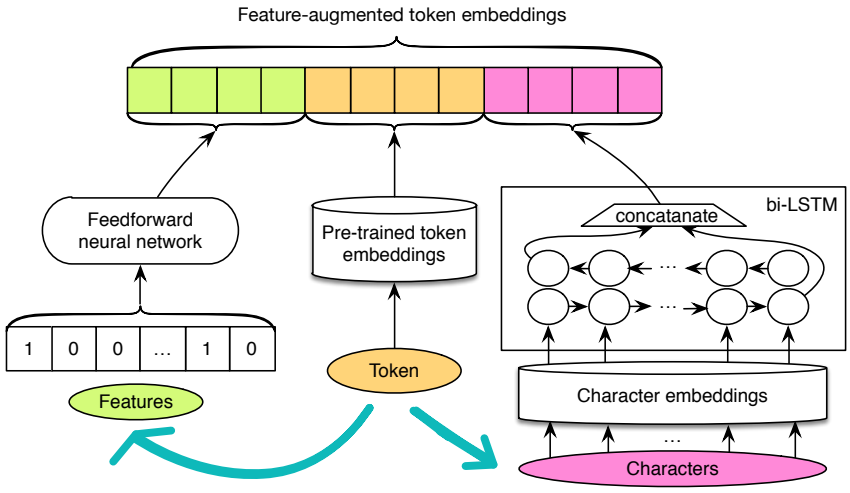

Figure 1: Feature-augmented token embeddings. Each token is mapped to a token embedding that is the concatenation of three elements: the output of a feedforward neural network that takes the features as input, a pre-trained token embedding, and the output of a bidirectional-LSTM (bi-LSTM) that takes the character embeddings as input.

Feature types	Features	
Note metadata	Patient's first name, patient's last name	EHR features
Hospital data	Doctor's first names, doctor's last names	
Morphological	Ends with s, is the first letter capitalized, contains a digit, is numeric, is alphabetic, is alphanumeric, is title case, is all lower case, is all upper case, is a stop word	
Semantic/Wordnet	Hypernyms, senses, lemma names	
Temporal	Seasons, months, weekdays, times of the day, years, years followed by apostrophe, festivity dates, holidays, cardinal numbers, decades, fuzzy quantifier (e.g., "approximately", "few"), future trigger (e.g., "next", "tomorrow")	
Gazetteers	Honorifics for doctors, honorifics, medical specialists, medical specialties, first names, last names, last name prefixes, street suffixes, US cities, US states (including abbreviations), countries, nationalities, organizations, professions	
Regular expressions	Email, age, date, phone, zip code, id number, medical record number	

Table 1: Feature list. Note metadata and hospital data are derived from the EHR database. Morphological, semantic/wordnet, and temporal features are commonly used features for NLP tasks. Gazetteers and regular expressions are specifically engineered for the task.

3 Experiments

We evaluate our model on the de-identification dataset introduced in (Dernoncourt et al., 2016), which is a subset of the MIMIC-III dataset (Goldberger et al., 2000; Saeed et al., 2011; Johnson et al., 2016), using the same train/validation/test split (70%/10%/20%). We chose this dataset as each note comes with metadata, such as the patient's name, and it is the largest de-identification dataset available to us. It contains 1,635 discharge summaries, 2,945,228 tokens, 69,525 unique tokens, and 78,633 PHI tokens.

The model is trained using stochastic gradient descent, updating all parameters, i.e., token embeddings, character embeddings, parameters of bidirectional LSTMs, and transition probabilities, at each gradient step. For regularization, dropout is applied to the character-enhanced token embeddings before the label prediction layer. We set the character embedding dimension to 25, the character-based token embedding LSTM dimension to 25, the token embedding dimension to 100, the label prediction LSTM dimension to 100, the dropout probability to 0.5, and we use GloVe embeddings (Pennington et al., 2014) trained on Wikipedia and Gigaword 5 (Parker et al., 2011) articles as pre-trained token embeddings. The hyperparameters were optimized based on the performance on the validation set.

4 Results

Table 2 presents the main results. The epochs for which the results are reported are optimized based on either the highest F1-score or the highest recall on the validation set. As expected, choosing the epoch based on the recall improves the recall on the test set, while lowering the precision. Overall, adding features consistently improves the results.

Table 3 details the results for each PHI type. The system using only the EHR features yields the highest recall for 6 out of 12 PHI types. Most importantly, the recall for patient and doctor names are higher when using features than when using no feature: this is expected as the patient name of the note and the doctor names are used as features. In fact, the two remaining false negatives for patient names are annotation errors. For example, in the sentence "The patient responded well to *Natrecor* in the past, but the improvement disappeared soon", the drug name *Natrecor* was incorrectly marked as a patient name by the human annotator. This result is highly remarkable as patient names are the most sensitive information in a patient note (South et al., 2014).

Adding all features often lowers the recall compared to using EHR features only, although the F1-score remains virtually unchanged. This is somewhat surprising, as we had expected that the features would help, as using the same feature set with a CRF to perform de-identification yields state-of-the-art results next to the ANN models (Dernoncourt et al., 2016). This could be explained as follows. Human-engineered features tend to have higher precision than recall, as it is often hard to design regular expressions or gazetteers that can detect all possible instances or variations of the desired entities. We

	Binary HIPAA (optimized by F1-score)			Binary HIPAA (optimized by recall)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
No feature	99.103	99.197	99.150	98.557	99.376	98.965
EHR features	99.100	99.304	99.202	98.771	**99.441**	99.105
All features	**99.213**	**99.306**	**99.259**	**98.880**	99.420	**99.149**

Table 2: Binary HIPAA token-based results (%) for the ANN model, averaged over 5 runs. The metric refers to the detection of PHI tokens versus non-PHI tokens, amongst PHI types that are defined by HIPAA. "No feature" is the model utilizing only character and word embeddings, without any feature. "EHR features" uses only 4 features derived from EHR database: patient first name, patient last name, doctor first name, and doctor last name. "All features" makes use of all features, including the EHR features as well as other engineered features listed in Table 1. "Optimized by F1-score" and "optimized by recall" means that the epochs for which the results are reported are optimized based on the highest F1-score or the highest recall on the validation set, respectively.

	No feature			EHR features			All features			
	P	R	F1	P	R	F1	P	R	F1	Support
Zip	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	24
Date	98.90	99.77	99.33	98.95	**99.79**	**99.36**	**98.99**	99.69	99.34	20627
Phone	98.31	**99.58**	98.94	**98.98**	99.46	99.22	99.42	99.32	**99.37**	1438
Patient	96.89	98.34	97.61	98.62	99.14	98.88	**99.21**	**99.27**	**99.24**	302
ID	99.57	98.24	98.90	99.31	**98.82**	**99.07**	**99.77**	97.97	98.86	612
Doctor[1]	97.47	98.17	97.82	97.27	**98.48**	97.87	**97.56**	98.20	**97.88**	3676
Location	96.02	95.71	95.86	96.41	**96.49**	96.45	**96.65**	96.32	**96.46**	462
Age ≥ 90	75.12	94.29	83.60	77.04	**95.72**	**85.35**	**78.93**	93.57	84.80	28
Hospital[1]	94.78	95.39	95.08	94.77	**95.52**	95.14	**95.53**	95.50	**95.51**	1259
State[1]	99.36	**94.33**	**96.76**	**99.68**	94.03	96.73	99.39	91.94	95.49	67
Street	96.77	85.25	90.54	**97.63**	85.25	**90.96**	93.91	**86.56**	89.81	61
Country[1]	87.51	85.00	86.11	**89.29**	82.50	85.67	86.87	**95.00**	**90.56**	16
Binary	98.41	99.19	98.80	98.48	**99.27**	98.87	**98.61**	99.15	**98.88**	28572

Table 3: Binary token-based results (%) . The reported results are optimized by recall, and averaged over 5 runs. The symbol [1] indicates that the PHI type is not required by HIPAA. The PHI type "location" designates any location that is not a street name, zip code, state or country. P stands for precision, R for recall, and F1 for F1-score.

conjecture that as the ANN model learn to rely more on such features, it might lose the ability to learn to pick up tokens that deviate from engineered features, resulting in a lower recall. For example, we notice that the phone PHI tokens that are not detected by the model using all features but are detected by the other two models, are ill-formed phone numbers such as "617-554-|2395", or phone extensions such as "617-690-4031 ext 6599". Since the phone regular expressions do not capture these two examples, they are more likely to be false negatives in the model that uses the phone regular expression features.

5 Conclusion

In this paper we presented an extension of the ANN-based model for patient note de-identification that can incorporate features. We showed that adding features results in an increase of the recall, in particular features leveraging information from the associated EHRs, namely patient names and doctor names.

Our results suggest that constructing patient note de-identification systems should be performed using structured information from the EHRs, the latter being available in a typical, real-life setting. We restricted our EHR-derived features to patient and doctor names, but it could be extended to the many other structured fields that EHR contain, such as patients' addresses, phone numbers, email addresses, professions, and ages.

Acknowledgements

The project was supported by Philips Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of Philips Research. We warmly thank Michele Filannino, Alistair Johnson, Li-wei Lehman, Roger Mark, and Tom Pollard for their helpful suggestions and technical assistance.

References

[Aberdeen et al.2010] John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. 2010. The MITRE Identification Scrubber Toolkit: design, training, and assessment. *International journal of medical informatics*, 79(12):849–859.

[Beckwith et al.2006] Bruce A Beckwith, Rajeshwarri Mahaadevan, Ulysses J Balis, and Frank Kuo. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC medical informatics and decision making*, 6(1):1.

[Berman2003] Jules J Berman. 2003. Concept-match medical data scrubbing: how pathology text can be used in research. *Archives of pathology & laboratory medicine*, 127(6):680–686.

[Dernoncourt et al.2016] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *arXiv preprint arXiv:1606.03475*.

[DesRoches et al.2013] Catherine M DesRoches, Chantal Worzala, and Scott Bates. 2013. Some hospitals are falling behind in meeting meaningful use criteria and could be vulnerable to penalties in 2015. *Health Affairs*, 32(8):1355–1360.

[Douglas et al.2004] Margaret Douglas, Gari Clifford, Andrew Reisner, George Moody, and Roger Mark. 2004. Computer-assisted de-identification of free text in the mimic ii database. In *Computers in Cardiology, 2004*, pages 341–344. IEEE.

[Douglass et al.2005] Margaret Douglass, Gari Cliffford, Andrew Reisner, William Long, George Moody, and Roger Mark. 2005. De-identification algorithm for free-text nursing notes. In *Computers in Cardiology, 2005*, pages 331–334. IEEE.

[Fielstein et al.2004] Elliot M. Fielstein, Steven H. Brown, and Theodore Speroff. 2004. Algorithmic de-identification of VA medical exam text for HIPAA privacy compliance: Preliminary findings. *Medinfo*, 1590.

[Filannino and Nenadic2015] Michele Filannino and Goran Nenadic. 2015. Temporal expression extraction with extensive feature type selection and a posteriori label adjustment. *Data & Knowledge Engineering*, 100:19–33.

[Friedlin and McDonald2008] Jeff Friedlin and Clement J McDonald. 2008. A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, 15(5):601–610.

[Goldberger et al.2000] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

[Guo et al.2006] Yikun Guo, Robert Gaizauskas, Ian Roberts, George Demetriou, and Mark Hepple. 2006. Identifying personal health information using support vector machines. In *i2b2 workshop on challenges in natural language processing for clinical data*, pages 10–11.

[Gupta et al.2004] Dilip Gupta, Melissa Saul, and John 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(2):176–186.

[Hara2006] Kazuo Hara. 2006. Applying a SVM based chunker and a text classifier to the deid challenge. In *i2b2 Workshop on challenges in natural language processing for clinical data*, pages 10–11. Am Med Inform Assoc.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hsiao et al.2011] Chun-Ju Hsiao, Esther Hing, Thomas C Socey, and Bill Cai. 2011. Electronic health record systems and intent to apply for meaningful use incentives among office-based physician practices: United states, 2001–2011. *system*, 18(17.3):17–3.

[Johnson et al.2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3.

[McCann2015] Erin McCann. 2015. EHR vendor marketshare and MU attestations by vendor. *Healthcare IT News*.

[Meystre et al.2010] Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1.

[Morrison et al.2009] Frances P Morrison, Li Li, Albert M Lai, and George Hripcsak. 2009. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? *Journal of the American Medical Informatics Association*, 16(1):37–39.

[Neamatullah et al.2008] Ishna Neamatullah, Margaret Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):1.

[Office for Civil Rights2002] HHS Office for Civil Rights. 2002. Standards for privacy of individually identifiable health information. final rule. *Federal Register*, 67(157):53181.

[Parker et al.2011] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition, linguistic data consortium. Technical report, Technical Report. Linguistic Data Consortium, Philadelphia.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

[Ruch et al.2000] Patrick Ruch, Robert H Baud, Anne-Marie Rassinoux, Pierrette Bouillon, and Gilbert Robert. 2000. Medical document anonymization with a semantic lexicon. In *Proceedings of the AMIA Symposium*, page 729. American Medical Informatics Association.

[Saeed et al.2011] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.

[South et al.2014] Brett R South, Danielle Mowery, Ying Suo, Jianwei Leng, Óscar Ferrández, Stephane M Meystre, and Wendy W Chapman. 2014. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. *Journal of biomedical informatics*, 50:162–172.

[Stubbs et al.2015] Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

[Sweeney1996] Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the Scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association.

[Szarvas et al.2006] György Szarvas, Richárd Farkas, and András Kocsor. 2006. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In *Discovery Science*, pages 267–278. Springer.

[Thomas et al.2002] Sean M Thomas, Burke Mamlin, Gunther Schadow, and Clement McDonald. 2002. A successful technique for removing names in pathology reports using an augmented search and replace method. In *Proceedings of the AMIA Symposium*, page 777. American Medical Informatics Association.

[Uzuner et al.2008] Özlem Uzuner, Tawanda C Sibanda, Yuan Luo, and Peter Szolovits. 2008. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, 42(1):13–35.

[Wright et al.2013] Adam Wright, Stanislav Henkin, Joshua Feblowitz, Allison B McCoy, David W Bates, and Dean F Sittig. 2013. Early results of the meaningful use program for electronic health records. *New England Journal of Medicine*, 368(8):779–780.

[Yang and Garibaldi2015] Hui Yang and Jonathan M Garibaldi. 2015. Automatic detection of protected health information from clinic narratives. *Journal of biomedical informatics*, 58:S30–S38.