Journal of Biomedical Informatics 75 (2017) S28–S33

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Learning to identify Protected Health Information by integrating knowledge- and data-driven algorithms: A case study on psychiatric evaluation notes

Azad Dehghan [a,b], Aleksandar Kovacevic [c], George Karystianis [d], John A Keane [a,f], Goran Nenadic [a,e,f,g,*]

[a] School of Computer Science, University of Manchester, Manchester, UK
[b] The Christie NHS Foundation Trust, Manchester, UK
[c] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[d] Macquarie University, Australian Institute of Health Innovation, Australia
[e] Health eResearch Centre, The Farr Institute of Health Informatics Research, UK
[f] Manchester Institute of Biotechnology, Manchester, UK
[g] Mathematical Institute, SANU, Serbia

ARTICLE INFO

Article history:
Received 2 February 2017
Revised 1 June 2017
Accepted 5 June 2017
Available online 7 June 2017

Keywords:
De-identification
Named entity recognition
Information extraction
Clinical text mining
Electronic health record

ABSTRACT

De-identification of clinical narratives is one of the main obstacles to making healthcare free text available for research. In this paper we describe our experience in expanding and tailoring two existing tools as part of the 2016 CEGS N-GRID Shared Tasks Track 1, which evaluated de-identification methods on a set of psychiatric evaluation notes for up to 25 different types of Protected Health Information (PHI). The methods we used rely on machine learning on either a large or small feature space, with additional strategies, including two-pass tagging and multi-class models, which both proved to be beneficial. The results show that the integration of the proposed methods can identify Health Information Portability and Accountability Act (HIPAA) defined PHIs with overall F_1-scores of ~90% and above. Yet, some classes (*Profession*, *Organization*) proved again to be challenging given the variability of expressions used to reference given information.

1. Introduction

Clinical free text data (including, for example, consultation notes, discharge letters, imaging reports etc.) contain a number of variables that are key for understanding patients' health conditions and their responses to treatments. Extracting such information is challenging due to inherent ambiguity and variability of clinical text, but one of the main obstacles to accessing such data in the first place is the presence of Protected Health Information (PHI). While de-identification and pseudo-anonymization of well-structured health data has been used routinely, it is still not clear what acceptable levels of masking PHI mentions in clinical narrative are [1–3].

The task of finding PHI instances in text is by and large a text mining task, where the aim is to identify mentions of specific PHI data types (e.g. patient names, age, address). This is a challenging task even for human annotators [4–6], and there have been several community challenges such as the 2006 i2b2 de-identification challenge [7], the 2014 i2b2/UTHealth Shared Task in de-identification of longitudinal clinical narratives [8]; with an increasing number of systems and papers addressing this issue [9]. The task is typically approached as named entity recognition (NER) of PHI data types. Two main approaches have been followed and quite often combined: knowledge-driven methods that rely on dictionaries and rules for regularized PHI types [10–13] and machine-learning and hybrid approaches that aim at learning from data [14–19]. The results of the community challenges have suggested that machine-learning approaches, in principle, provide better and more consistent performance [7,20].

A recent challenge in this area (the 2016 CEGS N-GRID Shared Tasks Track 1b [21]) further focused on NER of up to 25 PHI types (see Table 1). The organizers provided a high-quality training and a held-out test data set of initial psychiatric evaluation notes. In this paper we describe two methods developed and evaluated as part of that task, as well as the outcome of their integration. Our methods rely on previous work [22]. mDEID is a knowledge-driven approach

* Corresponding author at: School of Computer Science, University of Manchester, Manchester, UK.
E-mail addresses: azad.dehghan@manchester.ac.uk (A. Dehghan), kocha78@uns.ac.rs (A. Kovacevic), george.karystianis@mq.edu.au (G. Karystianis), john.keane@manchester.ac.uk (J.A Keane), goran.nenadic@manchester.ac.uk (G. Nenadic).

http://dx.doi.org/10.1016/j.jbi.2017.06.005
1532-0464/

Table 1
Composition of the submissions. CRF(mDEID) denotes the CRF-expanded version of mDEID; all references to CliDEID refer to the new version introduced here. Count is the number of instances in the held-out data; Union represents merging of the results as explained below.

Entity type	COUNT	Submission1	Submission 2	Submission 3
Date	3822	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Age	2354	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Doctor	1567	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Hospital	1328	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Profession	1010	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Patient	837	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
City	820	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Organization	697	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
Country	376	Union(Sub2,Sub3)	CRF(mDEID)	CliDEID
State	481	mDEID	mDEID	mDEID
Phone	113	mDEID	mDEID	mDEID
Street	34	mDEID	mDEID	mDEID
License	21	mDEID	mDEID	mDEID
Zip	17	mDEID	mDEID	mDEID
Idnum	8	mDEID	mDEID	mDEID
Email	5	mDEID	mDEID	mDEID
Fax	5	mDEID	mDEID	mDEID
Url	3	mDEID	mDEID	mDEID

that relies on dictionaries to identify relatively closed PHI types (e.g. *Country*, *State*) and a generic set of lexico-syntactic rules that model common orthographic and contextual characteristic of specific PHI types (e.g. Addresses, Phone numbers). On the other hand, CliDEID is a CRF-based tagger that uses 279 features grouped into lexical, orthographic, semantic and positional attributes. In this paper we build on top of these two approaches by adding a learning Conditional Random Fields (CRF) layer on top of mDEID and introducing multi-class labeling into CliDEID. One of our key aims was to explore how re-usable existing de-identification methods are when migrated to new settings (e.g. a move from cancer discharge notes to psychiatric evaluation notes). The results (with an overall HIPAA strict F_1 score of ~90%, ranking our system within top 3) show the potential and challenges introduced by both data-driven methods with rich (large) and focused (small) feature sets, as well as the benefits of additional processing, including two-pass tagging, multi-class models, and label priority sorting.

The following section explains the details of the proposed methodology. Section 3 presents the results and discussion, which are followed by the conclusion.

2. Method

The approaches we designed are built using two previously published methods [22], which include a knowledge-driven open source algorithm (mDEID) and a data-driven method (CliDEID) built using linear chain CRF. We used default (CRF++) parameters: L2-regularization with C = 1.00, ETA = 0.001. For some PHI types, mDEID was expanded by providing an additional CRF layer that mainly relies on rules and dictionaries as features. CliDEID on the other hand was expanded by training models for multi-class labeling for a selected set of PHI types. We have submitted three versions for official evaluation: Submission 1 combined the outputs of Submission 2 (based on mDEID) and Submission 3 (mainly based on CliDEID). Table 1 provides the details, which are further explained below.

Submission 2 is built on top of mDEID, which was initially modeled on the i2b2/UTHealth 2014 Track I [22,23]. The rules already available in mDEID were updated based on the new training data. Further, six additional NER components were developed for *Date, Hospital, Profession, City, Organization* and *License*. In addition, CRF models were trained for nine categories using a small and focused set of features generated by the mDEID pipeline. The **B**eginning-**I**nside-**O**ut (BIO) token representation was used. The core set of features used include (see Supplement, Appendix B for per category feature set):

- *Lexical features*, such as the word/token, its stem (derived from Porter's stemmer), part-of-speech and shallow parsing information.
- *Orthographic features*, including token characteristics such as word casing (upper initial, all capital, lower case, and mixed capitalization) and type (word, number, punctuation, and symbol).
- *Semantic features*, which are binary attributes indicating if a given token was tagged by mDEID knowledge-driven components.
- *Contextual features*, including a context window of two tokens before and two tokens after each current token.

We generated a minimum of 26 (*Age*) to a maximum of 44 (*Doctor*) features using forward and backward feature selection strategies. In addition, the two-pass recognition (see below) is adopted for a subset of entity types (*City*, *Country*, *Doctor*, *Hospital*, *Organization*, *Patient*, and *Profession*).

Submission 3 is a data-driven method developed on top of CliDEID, a machine learning component of our system developed for the 2014 de-identification challenge [22]. It relies on the same feature set (*lexical*, *orthographic*, *semantic*, *positional*) and the models were trained using the **I**nside-**O**utside (I-O) schema. Building on top of the 2014 system, CliDEID has the following newly introduced characteristics:

- Models with multiple class labels. In contrast to the previous version where each CRF model was aimed at a specific category and trained only with the class labels of that particular category, a subset of the CliDEID models was trained with multiple category class labels. This was done with the goals of (a) reducing confusions between lexically similar categories (e.g. 'George' can be either a *Patient* or a *Doctor*; 'Harvard' can be either a *City* or *Hospital* or an *Organization*) and (b) exploiting the fact that some of the categories frequently occur in a sequence in the same sentence (e.g. *Patient* and *Age* – 'Valentina is a 43-year old' or *Profession* and *Organization* – 'Works as medical assistant at MEDIQUIK'). We created five multi-label machine learning (ML) models: (1) *Age* and *Patient*, (2) *City*, *Doctor*, *Hospital*, *Patient* and *Organization*, (3) *Patient* and *Doctor*, and (4–5) two models for *Organization* and *Profession*, one optimized for each of the two classes. Each of the models generates separate labels

for each of the classes it models. In addition, single-class label models were trained for *Country*, *Date*, *Doctor*, and *Hospital* categories.

- Combining multiple label outputs. As a result of the multiple class labeling step, the system can produce multiple tags for the *Patient*, *Doctor*, *Hospital, Professional* and *Organization* categories (e.g., three models produced tags for the *Doctor* category, see Fig. 1). Based on the results on the validation set, we used a union of all the *Doctor, Organization,* and *Profession* entities produced by the corresponding models as the output of the system. For example, each of the three CRF models shown in Fig. 1 outputs (amongst other categories) the tags for the *Doctor* category; the final output for the *Doctor* category is the union of all the entities annotated by *Doctor* tags produced by the three CRF outputs. However, for the *Patient* and *Hospital* categories, the validation data has not supported adding the results from the multiple-class taggers to the corresponding single-class models, and thus we only used the output of the single-class models for these classes.
- Using additional training data. Our results in a similar clinical NER challenge in 2012 [24] showed that using supplementary training data (in addition to the one provided by the organizers) could have a positive effect on the performance of the ML models. Based on that, we experimented with enriching the training data with the data set provided in the 2014 de-identification challenge. Following the validation results, we decided to add the 2014 data to the training sets for the *Doctor* and *Hospital* models.
- Bigram features. The CRF models were trained using the CRF++ software [25] that enables automatic generation of bigram features (combinations of feature values for the current token and the previous one - bigram). After experiments on the validation set (improvements in precision with a very slight drop in recall) we opted to include the bigram features for *Age*, *City*, *Country*, *Doctor, Hospital*, *Organization*, *Profession* and *Patient* models.

As a final step, the CliDEID system used the 'two-pass tagging' and 'priority sorting' approaches (see below) to produce additional tags and resolve conflicts arising from multiple models tagging the same text span.

As indicated above, Submissions 2 and 3 use additional processing after the main steps are done. The **two-pass tagging** method, previously shown effective on longitudinal clinical narratives (at the patient-level processing), uses the outputs of initial NER steps to generate document-specific dictionaries that were propagated on the same document (i.e., document-level processing) by dictionary matching. We have also used **priority sorting** for conflicting annotations where a given span could belong to more than one category. For example, in the following sentence '*Stepbrother is in Delaware.*', '*Delaware*' is tagged both as *City* and *State* by our models. In order to resolve the conflict, we gave a higher priority to the *State* model as it provided better performance on the validation set (and thus as the final output we produced only the *State* tag).

Finally, Submission 1 integrates the outputs of Submissions 2 and 3 through the union of the results at the entity-level for the categories where Submissions 2 and 3 had separate outputs; in cases of tag overlap, only the longer span was kept. The remaining entity types were adopted from Submission 2.

3. Results and discussion

The methods presented above were trained on 600 notes and tested on a held-out data set of 400 notes. Submission 1 showed the best performance across different evaluation settings (see Table 2) including the strict micro F_1-score of 87.69%.

The results for Submission 1 were expected as our findings on the validation results (data not shown) indicated that the two systems produce slightly different sets of tags that complement each other in their union. For example, Table 3 shows the F1 agreement (derived by considering one system as gold and the other as predictions) between mDEID and CliDEID compared to the results of Submission 1.

Submission 1 uses the union at the entity-level by combining entities tagged by both systems. During the validation (results not shown) we experimented with different system integration configurations. For example, Table 4 shows merging the outputs by two votes on exact boundary match (intersection), which opti-

Table 2
Evaluation results (strict F_1-score %) on the held-out data set.

	Submission 1	Submission 2	Submission 3
All token	90.97	89.55	88.57
All strict	87.69	85.65	85.72
All relaxed	88.13	86.21	85.93
HIPAA token	92.73	91.89	90.48
HIPAA strict	89.93	88.39	87.71
HIPAA relaxed	90.29	88.86	87.89

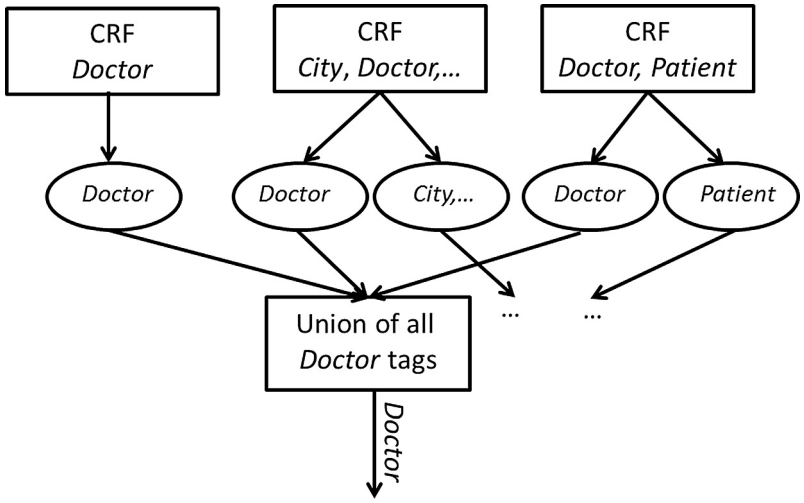

Fig. 1. An example of combining multiple CRFs producing output for the *Doctor* category.

Table 3
Per category agreement between mDEID and CliDEID compared with Submission 1. (On the held-out data set; all scores given in percentage).

Category	Agreement F_1-score	Submission 1 F_1-score	ΔF_1-score
Age	89.72	94.30	4.58
City	76.15	84.22	8.07
Country	75.91	81.23	5.32
Date	94.42	94.73	0.31
Doctor	88.29	92.84	4.55
Hospital	80.32	83.08	2.76
Organization	55.75	56.20	0.45
Patient	82.38	86.59	4.21
Profession	64.33	69.25	4.92
Micro	85.48	87.76	2.28

mized precision (97.50% vs 88.79%), whilst union at the entity-level (merging by at least one vote) optimized F_1-score (87.76% vs 83.67%).

As highlighted in previous work, the identification of *Organization* and *Profession* PHI entities is still the most challenging part, in particular with low recall values. This is the case across different submissions, as expression patterns of employment information are quite diverse. Indicative examples include (entities are underlined) *'and said "you're a good dancer, but the extra weight'* (*Profession*); *'but found medical classes too stressful'* (*Profession*); *'seeking work with support from Zenith Electronics'* (*Organization*). Of note is that dictionary (as well as on lexical and orthographic) features generated a subset of false positives (FPs) for these classes e.g. (false positives are underlined), *'may need coaching'* (*Profession*), *'living in condo in Worthington with husband, new construction'* (*Profession*), *'went to Narc Anon in Pennsylvania'* (*Organization*), *'instructions provided in your packet from GI Associates.'* (*Organization*), *'Harrington Rod placement'* (*City*), *'Boston Terrier'* (*City*). Also, in many cases the ML models produced partial identification of Organisation and Profession mentions e.g. (correctly tagged tokens are underlined), '*Computer and Information Systems Manager*' (*Profession*), '*vaughan bushnell manufacturing*' (*Organization*) etc.

An interesting observation between Submission 2 and Submission 3 is that the latter performed generally better on the strict and the former on relaxed metrics when considering entity-level evaluation (see Supplement, Table A.1). This is an interesting observation given the general differences between the pipelines. For instance, Submission 2 was built using a focused feature set (26–43 features across trained models) while Submission 3 was trained using a rich features set (~460 features) with focused dictionaries used for five NERs (*Country*, *Profession*, *Hospital*, *City*, *Organization*). Hence, this may indicate that rich feature sets can help boundary identification at the cost of recall.

In our previous work we proposed and validated a two-pass tagging method for identifying PHI in longitudinal clinical narratives. This method is similar to [16,26], with the difference that (a) "trusted term lists" are generated by including all mentions tagged (except for ambiguous terms identified during training/development and subsequently filtered) by specific NERs in the first pass, and (b) the resulting entity-specific dictionaries and dictionary matching in the second pass were used as final output. We note that our method was developed independently of other studies using similar approaches. We investigated two-pass tagging on non-longitudinal records and found that this method was also equally useful (see Table 5 versus Supplement, Table A.1). For instance, two-pass tagging yielded improvements in F1-score on the held-out data set for Submission 1 (~1%), Submission 2 (~1.5%), and Submission 3 (~1.5%). Priority sorting also proved

Table 4
Integration of Submission 2 and Submission 3 pipelines. (On the held-out data; all scores given in percentage).

	Merging by at least one vote (Submission 1)			Merging by two votes (intersection)		
Category	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Age	95.47	93.16	94.30	98.89	79.35	88.05
City	80.60	88.17	84.22	96.99	66.83	79.13
Country	78.13	84.57	81.23	97.19	64.36	77.44
Date	95.21	94.24	94.73	98.12	89.01	93.34
Doctor	90.64	95.15	92.84	99.15	82.26	89.92
Hospital	84.61	81.61	83.08	93.90	63.83	76.00
Organization	64.04	50.07	56.20	90.64	30.56	45.71
Patient	90.60	82.92	86.59	98.01	64.87	78.07
Profession	73.66	65.35	69.25	93.15	43.07	58.90
Micro	88.79	86.75	87.76	97.50	73.28	83.67

Table 5
Strict per category results on the held-out data set – No Two-Pass Tagging. (All scores given in percentage).

	Submission 1			Submission 2			Submission 3		
Category	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Age[a]	95.51	93.16	94.32	95.77	90.36	92.98	97.64	82.54	89.46
City	79.02	86.34	82.52	89.34	70.49	78.80	80.83	80.73	80.78
Country	85.99	81.65	83.77	87.21	70.74	78.12	93.84	68.88	79.45
Date[a]	95.21	94.24	94.73	95.79	92.31	94.02	96.04	91.99	93.97
Doctor	88.84	94.00	91.35	92.99	83.79	88.15	93.83	90.24	92.00
Hospital	87.35	74.38	80.34	86.49	66.09	74.93	90.88	63.07	74.47
Organization	61.90	46.63	53.19	68.75	37.88	48.84	65.96	40.03	49.82
Patient	93.83	65.35	77.04	92.76	50.54	65.43	96.52	56.27	71.09
Profession	73.14	65.25	68.97	79.57	54.75	64.87	75.13	58.02	65.47
Micro	89.05	84.32	86.62	91.79	77.51	84.05	91.73	77.80	84.20

[a] Two Pass Tagging was not used at any-point for Submission 2.

Table 6
The difference in performance of the CliDEID system with new characteristics on the held-out data. (All scores given in percentage).

	CliDEID			CliDEID without multi-class label				CliDEID without combinations			
Category	Precision	Recall	F_1-score	Precision	Recall	F_1-score	ΔF_1-score	Precision	Recall	F_1-score	ΔF_1-score
Age	97.59	82.50	89.41	97.58	82.07	89.16	−0.25	N/A			
City	83.33	80.49	81.89	84.48	76.34	80.20	−1.69	N/A			
Doctor	94.71	90.24	92.42	96.90	83.79	89.87	−2.55	97.11	83.66	89.89	−2.53
Organization	74.81	42.18	53.94	81.27	34.86	48.80	−5.14	71.94	37.16	49.01	−4.93
Patient	89.84	75.03	81.77	90.73	72.52	80.61	−1.16	N/A			
Profession	76.11	57.72	65.65	80.53	51.19	62.59	−3.06	81.44	52.57	63.90	−1.75

useful. For example, CliDEID showed improvement of micro F_1 of ∼1.5% and mDEID around ∼2% on the held-out data set.

However, the two-pass processing can also propagate false positives (FPs). For example, in Submission 3, around 14% of FPs came from the FP propagation. The mDEID system used a more sophisticated second pass system which reduced the percentage of repeated false negatives (FNs) by ∼3% in Submission 1, and only added ∼6% to the repeated FPs (the false positives could only be increased in this submission as it was a union of the two system outputs).

We found that models with multi-label outputs are useful for capturing entities that commonly co-occur in text. We used the validation data set to assess that impact, which was largely replicated on the held-out test data (Table 6). Both the multi-label training approach and the combination of multiple label outputs had an overall larger positive effect on recall then negative on precision, which has led to improvements in F_1-score, including ∼5% for *Organization*, and ∼3% for *Profession*.

4. Conclusion

In this paper we presented two approaches to identify PHI in clinical text. We expanded existing methods by adding additional learning features and then combined the outcomes. Although not with a huge margin, the combined output provided the best performance. We have shown that the two-pass approach, initially proposed for longitudinal records is also beneficial for non-longitudinal data sets, as was the multi-label models and priority sorting. Although generalization of de-identification NER methods can be challenging on different data sets (i.e., different hospitals and clinical domain), we have also shown that methods modeled on different data can be reused through rapid development and re-training with very good performance. Further work still need to focus on improving the identification of classes where the recall is low – in particular *Profession* and *Organization* entity types by exploring unsupervised approached as well as common methods described in [27].

Availability

The mDEID software is available as open source at www.clinical-deid.sourceforge.net. The CliDEID software is available at www.github.com/kovacevica/CliDEID.

Conflict of interest

We declare no conflict of interest..

Acknowledgments

GN was partially supported by the UK's Farr Institute of the Health Informatics Research, Health eResearch Centre, and AK, GN by the Serbian Ministry of Education and Science (projects III44006; III47003). We also acknowledge the following grants: NIH P50 MH106933; NIH 4R13LM011411.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.06.005.

References

[1] S.M. Meystre, Ó. Ferrández, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Text de-identification for privacy protection: a study of its impact on clinical text information content, J. Biomed. Inform. 50 (2014 Aug) 142–150, http://dx.doi.org/10.1016/j.jbi.2014.01.011. PubMed PMID: 24502938.
[2] D. Carrell, B. Malin, J. Aberdeen, S. Bayer, C. Clark, B. Wellner, et al., Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text, J. Am. Med. Inform. Assoc. 20 (2) (2013) 342–348, http://dx.doi.org/10.1136/amiajnl-2012-001034.
[3] M. Scaiano, G. Middleton, L. Arbuckle, V. Kolhatkar, L. Peyton, M. Dowling, et al., A unified framework for evaluating the risk of re-identification of text de-identification tools, J. Biomed. Inform. 63 (2016) 174–183, http://dx.doi.org/10.1016/j.jbi.2016.07.015.
[4] M. Kayaalp, A.C. Browne, P. Sagan, T. McGee, C.J. McDonald, Challenges and insights in using HIPAA privacy rule for clinical text annotation, in: Proceedings of the AMIA Annual Symposium, Chicago, IL, 2015, pp. 707–716.
[5] D.S. Carrell, D.J. Cronkite, B.A. Malin, J.S. Aberdeen, L. Hirschman, Is the Juice Worth the Squeeze? Costs and benefits of multiple human annotators for clinical text de-identification, Methods Inf. Med. 55 (4) (2016) 356–364, http://dx.doi.org/10.3414/ME15-01-0122.
[6] B.R. South, D. Mowery, Y. Suo, J. Leng, O. Ferrandez, S.M. Meystre, et al., Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text, J. Biomed. Inform. 50 (2014) 162–172, http://dx.doi.org/10.1016/j.jbi.2014.05.002.
[7] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, JAMIA 14 (5) (2007) 550–563, http://dx.doi.org/10.1197/jamia.M2444.
[8] A. Stubbs, C. Kotfila, O. Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1, J. Biomed. Inform. 58 (Suppl) (2015 Dec) S11–S19, http://dx.doi.org/10.1016/j.jbi.2015.06.007. Review. PubMed PMID: 26225918; PubMed Central PMCID: PMC4989908.
[9] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Med. Res. Methodol. 2 (10) (2010 Aug) 70, http://dx.doi.org/10.1186/1471-2288-10-70. Review. PubMed PMID: 20678228; PubMed Central PMCID: PMC2923159.
[10] I. Neamatullah, M.M. Douglass, L.-W.H. Lehman, et al., Automated de-identification of free-text medical records, BMC Med. Inform. Decis. Mak. 8 (32) (2008). https://dx.doi.org/10.1186/1472-6947-8-32.
[11] F.P. Morrison, A.M. Lai, G. Hripcsak, Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes?, JAMIA 16 (1) (2009) 37–39, http://dxdoi.org/10.1197/jamia.M2862.
[12] A.C. Fernandes, D. Cloete, M.T. Broadbent, et al., Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records, BMC Med. Inform. Decis. Mak. 13 (71) (2013). https://dx.doi.org/10.1186/1472-6947-13-71.
[13] R. Guillen, Automated de-identification and categorization of medical records, in: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
[14] E. Aramaki, T. Imai, K. Miyo, K. Ohe, Automatic deidentification by using sentence features and label consistency, in: Paper Presented at: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
[15] Y. Guo, R. Gaizauskas, I. Roberts, G. Demetriou, M. Hepple, Identifying personal health information using support vector machines, in: Paper Presented at: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.

[16] G. Szarvas, R. Farkas, R. Busa-Fekete, State-of-the-art anonymization of medical records using an iterative machine learning framework, J. Am. Med. Inform. Assoc. 14 (5) (2007) 574, http://dx.doi.org/10.1197/j.jamia.M2441.
[17] J. Gardner, L. Xiong, HIDE: An integrated system for health information DE-identification, in: Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems, 2008, pp. 254–259 https://dx.doi.org/10.1109/CBMS.2008.129.
[18] J. Aberdeen, S. Bayer, R. Yeniterzi, et al., The MITRE identification scrubber toolkit: design, training, and assessment, JAMIA 79 (12) (2010) 849–859, http://dx.doi.org/10.1016/j.ijmedinf.2010.09.007.
[19] Ö. Uzuner, T.C. Sibanda, Y. Luo, et al., A de-identifier for medical discharge summaries, Artif. Intell. Med. 42 (1) (2008) 13–35, http://dx.doi.org/10.1016/j.artmed.2007.10.001.
[20] A.C. Stubbs, Ö. Kotfila, Automated Uzuner, Systems for the de-identification of longitudinal clinical narratives: overview of, i2b2/UTHealth Shared Task Track 1 (2015), J. Biomed. Inform. 58S (2015) (2014) S11–S19.
[21] A. Stubbs, M. Filannino, O. Uzuner, De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1, J. Biomed. Inform. 75 (2017) S4–S18.
[22] A. Dehghan, A. Kovačević, G. Karystianis, J.A. Keane, G. Nenadic, Combining knowledge- and data-driven methods for de-identification of clinical narratives, J. Biomed. Inform. 58 (Supplement) (2015) S53–S59.
[23] A. Dehghan, T. Liptrot, D. Tibble, M. Barker-Hewitt, G. Nenadic, Identification of occupation mentions in clinical narratives, in: Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22–24, 2016. http://dx.doi.org/10.1007/978-3-319-41754-7_35.
[24] A. Kovacevic, A. Dehghan, M. Filannino, J. Keane, G. Nenadic, Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives, J. Am. Med. Inform. Assoc. http://dx.doi.org/10.1136/amiajnl-2013-00.
[25] T. Kudo, CRF++: Yet another crf toolkit (2005), 2005, Software available at https://taku910.github.io/crfpp/ (accessed 27.02.17).
[26] H. Yang, M.J. Garibaldi, Automatic detection of protected health information from clinic narratives, J. Biomed. Inform. 58 (2015) S30–S38.
[27] A. Dehghan, J.A. Keane, G. Nenadic, Challenges in clinical named entity recognition for decision support, in: 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, October 13-16, 2013. https://doi.org/10.1109/SMC.2013.166.