Building a De-identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text

Hanna Berg
Department of Computer
and Systems Sciences
Stockholm University
Kista, Sweden
hanna.berg@dsv.su.se

Taridzo Chomutare
Norwegian Centre for E-health Research
University Hospital of North Norway
Tromsø, Norway
Taridzo.Chomutare@ehealthresearch.no

Hercules Dalianis[*]
Department of Computer and Systems Sciences
Stockholm University
Kista, Sweden
hercules@dsv.su.se

Abstract

This article presents experiments with pseudonymised Swedish clinical text used as training data to de-identify real clinical text with the future aim to transfer non-sensitive training data to other hospitals.

Conditional Random Fields (CFR) and Long Short-Term Memory (LSTM) machine learning algorithms were used to train de-identification models. The two models were trained on pseudonymised data and evaluated on real data. For benchmarking, models were also trained on real data, and evaluated on real data as well as trained on pseudonymised data and evaluated on pseudonymised data.

CRF showed better performance for some PHI information like *Date Part, First Name* and *Last Name*; consistent with some reports in the literature. In contrast, poor performances on *Location* and *Health Care Unit* information were noted, partially due to the constrained vocabulary in the pseudonymised training data.

It is concluded that it is possible to train transferable models based on pseudonymised Swedish clinical data, but even small narrative and distributional variation could negatively impact performance.

1 Introduction

Electronic health records (EHR) are produced in a steady stream, with the potential of advancing future medical care. Research on EHR data holds the potential to improve our understanding of patient care, care processes, and disease characteristics and progression. However, much of the data is sensitive, containing Protected Health Information (PHI) such as personal names, addresses, phone numbers, that can identify particular individuals and thus cannot be available to the public for general scientific inquiry. Although good progress has been made in the general sub-field of de-identifying clinical text, the problem is still not fully resolved (Meystre et al., 2010; Yogarajan et al., 2018).

This study examines the use of pseudonymised health records as training data for de-identification tasks. Several ethical and scientific issues arise regarding the balance between maintaining patient confidentiality and the need for wider application of trained models. How will a de-identification system be constructed and used in a cross hospital setting without risking the privacy of patients? Is it possible to obscuring the training data by pseudonymising it and then use it for the training of a machine learning system?

De-identification and pseudonymisation are two related concepts. In this paper de-identification is used as a more general term to describe the process of finding personal health information to be able to conceal identifying information. A pseudonymised text is a text where the personal health information has been identified either manually or automatically and then replaced with realistic surrogates.

The research question in this study is whether it is possible to use de-identified and pseudonymised clinical text in Swedish as training data for de-identifying real clinical text, and hence make it possible to transfer the system cross hospital.

We highlight whether learning from the exist-

[*] Hercules Dalianis is also guest professor at the Norwegian Centre for E-health Research

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 118–125
Hong Kong, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/D19-62

ing, non-sensitive, pseudonymised Swedish clinical text can be useful in a new and different context; considering the normal variations in the distribution and nature of PHI information, and potential effects of scrubbing (Berman, 2003), that is, removing and modifying PHIs that was carried out to patient records during the de-identification process.

2 Previous research

The identification of PHI is a type of named entity recognition task where sensitive named entities specifically are identified. The first study with CRF-based de-identification for Swedish was on the gold standard Stockholm EPR PHI Corpus. The distribution of PHIs is shown in Table 1. In this instance, manual annotation with expert consensus was used to create the gold standard (Dalianis and Velupillai, 2010).

De-identification tasks based on the CRF machine learning algorithm has been carried out on this data set previously with precision scores ranging between 85% and 95%, recalls ranging between 71% and 87% and F1-scores between 0.76 and 0.91 (Dalianis and Velupillai, 2010; Berg and Dalianis, 2019).

One approach previously used for concealing the training set's sensitive data was carried out by (Dalianis and Boström, 2012), using the Stockholm EPR PHI Corpus. In the study, the textual part of the data were used to create 14 different features and part of speech tags. The textual part was then removed, and only the features and part of speech tags were used for training a Random Forest model. Fairly high precision of 89.1 % was obtained, but with a recall of 54.3 % and F1-score of 64.8.

In contrast to using only the sensitive EHR data for training, McMurry et al. (2013) integrated both publicly available scientific, medical publications and private sensitive clinical notes to develop a de-identification system. While considering the term frequencies and part of speech tags between the two data sources, they used both rule lists and decision trees for their system. This was an interesting approach since it raised the prospect of using non-sensitive data in building useful de-identification models. However, it is not clear whether medical journals have significant advantages over any other public text, like news corpora, for detecting PHI. A study similar to McMurry et al. (2013), by Berg and Dalianis (2019), showed few benefits of combining non-medical public text and sensitive clinical notes to build a de-identification system for medical records.

More recently, deep learning approaches using recurrent neural networks seem to yield significant improvements over traditional rules-based methods or statistical machine learning (Dernoncourt et al., 2017). Still, recent studies indicate that combining several approaches will yield the best results. For instance, the best system in a recent de-identification shared task was a combination of bidirectional LSTM, CRF and a rule-based subsystem (Liu et al., 2017).

Significant domain variation, such as a different language, is an important factor that was not considered in the discussed shared task. Domain differences were cited as the reason for poor performance on psychiatric notes de-identification (Stubbs et al., 2017), compared with the previous de-identification task on general clinical narratives (Stubbs et al., 2015).

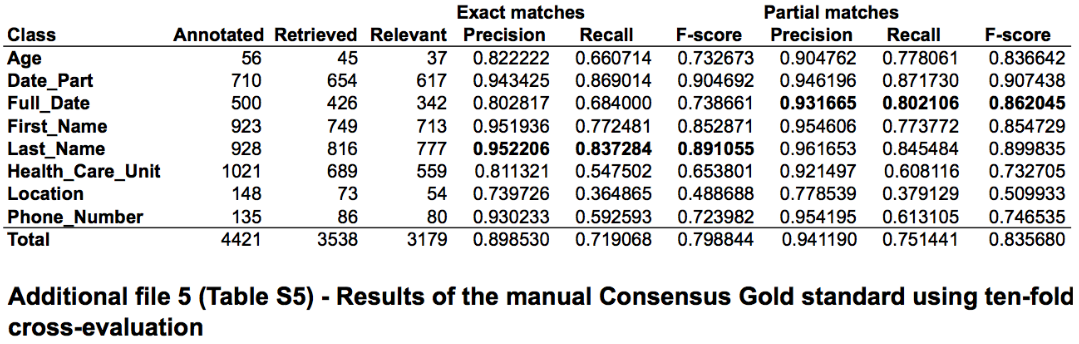
Within the same language and similar clinical settings, the change of domain is likely not substantial. While in future research it may be worth considering domain adaption techniques to work towards a system meant to be used between hospitals, they were not considered in this study, beyond the use of non-sensitive dictionaries for names and location.

3 Data and methods

In this study, machine learning approaches are used since the best de-identification systems appear to be machine learning-based (Kushida et al., 2012). While rule-based methods such as using dictionaries and pattern-matching were previously more prevalent than machine learning methods for solving text-based de-identification problems (Meystre et al., 2010), today it is more typical to have both approaches used, since rule-based methods still yield better results for some PHI information (Neamatullah et al., 2008b). Dictionaries and patterns were therefore used as features within one of the models.

3.1 Data

Two different data sets for de-identification were used: Stockholm EPR PHI Psuedo Corpus (*Pseudo*) as well as the Stockholm EPR PHI Cor-



Class	Annotated	Retrieved	Relevant	Exact matches Precision	Recall	F-score	Partial matches Precision	Recall	F-score
Age	56	45	37	0.822222	0.660714	0.732673	0.904762	0.778061	0.836642
Date_Part	710	654	617	0.943425	0.869014	0.904692	0.946196	0.871730	0.907438
Full_Date	500	426	342	0.802817	0.684000	0.738661	**0.931665**	**0.802106**	**0.862045**
First_Name	923	749	713	0.951936	0.772481	0.852871	0.954606	0.773772	0.854729
Last_Name	928	816	777	**0.952206**	**0.837284**	**0.891055**	0.961653	0.845484	0.899835
Health_Care_Unit	1021	689	559	0.811321	0.547502	0.653801	0.921497	0.608116	0.732705
Location	148	73	54	0.739726	0.364865	0.488688	0.778539	0.379129	0.509933
Phone_Number	135	86	80	0.930233	0.592593	0.723982	0.954195	0.613105	0.746535
Total	4421	3538	3179	0.898530	0.719068	0.798844	0.941190	0.751441	0.835680

Additional file 5 (Table S5) - Results of the manual Consensus Gold standard using ten-fold cross-evaluation

Table 1: Results from (Dalianis and Velupillai, 2010) using the Stanford CRF.

pus (*Real*)[1].

The Stockholm EPR PHI Pseudo Corpus was produced from the Stockholm EPR PHI Corpus by automatically pseudonymising all PHIs. This process is described by Dalianis (2019). The Stockholm EPR PHI Corpus is described by Dalianis and Velupillai (2010). An example is shown in Figure 1 from the Stockholm EPR PHI Pseudo Corpus.

The Stockholm EPR PHI Corpus and the Stockholm EPR PHI Pseudo Corpus are both parts of Swedish Health Record Research Bank (HEALTH BANK). HEALTH BANK encompasses structured and unstructured data from 512 clinical units from Karolinska University Hospital collected from 2006 to 2014 (Dalianis et al., 2015).

The number of entities and types of entities in both the Stockholm EPR PHI Psuedo Corpus and the Stockholm EPR PHI Corpus is shown in Table 2. From Table 2, it can be observed that the distribution of PHI instances between the two data sets is somewhat similar, but there is a significant difference when it comes to unique instances between the two data sets. In total, the *Real* data set contains proportionally more unique instances than the *Pseudo* data set. The entities in the *Real* data set also tend to have more tokens.

3.2 Methods

Using the de-identified and pseudonymised data set, two models were trained based on two machine learning algorithms; CRF and the deep learning algorithm LSTM. The two algorithms were chosen since both have been shown to produce state of the art performance, and applying the two on Swedish clinical data sets makes for an informative comparison.

PHI classes	Pseudo	Unique	Real	Unique
First Name	885	24 %	938	79 %
Last Name	911	15 %	957	86 %
Age	51	80 %	64	97 %
Phone Number	310	78 %	327	92 %
Location	159	94 %	229	84 %
Full Date	726	25 %	770	89 %
Date Part	1897	6 %	2079	72 %
Health Care Unit	1278	13 %	2277	73 %
Total PHI instances	6217	20 %	7647	78 %

Table 2: The distribution of PHI instances between the the Stockholm EPR PHI Psuedo Corpus, *'Pseudo'*, and the Stockholm EPR PHI Corpus, *'Real'* based on the number of tokens. A PHI entity can cover from one token (one-word expression) to several tokens (multi-word-expression), for example "Karolinska" and "R54, Karolinska, Solna" respectively. The proportion of unique instances, *'Unique'*, is shown as a percentage.

The two models were evaluated on both the real data set that is annotated for PHI, but not pseudonymised, *'Pseudo-Real'*, as well as on the pseudonymised data set, *'Pseudo-Pseudo'*. For additional comparison basis models trained on the real data set were evaluated on test sets from the same data set, *'Real-Real'*.

3.2.1 CRF

In this study, the CRF algorithm implemented in CRFSuite (Okazaki, 2007) is used with the sklearn-crfsuite wrapper[2] and the LSTM architecture described by Lample et al. (2016), based on an open-source implementation with Tensorflow[3] is used.

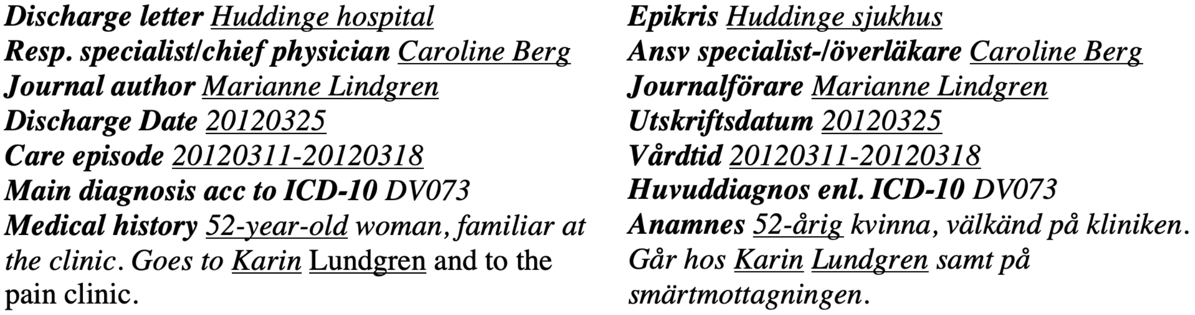
The linear-chain Conditional Random Fields model, implemented with sklearn-CRFSuite[4],

[1] This research has been approved by the Regional Ethical Review Board in Stockholm (2014/1607-32).

[2] sklearn-crfsuite, https://sklearn-crfsuite.readthedocs.io

[3] Sequence tagging, https://github.com/guillaumegenthial/sequence_tagging

[4] Linear-chain CRF, https://



Discharge letter *Huddinge hospital*
Resp. specialist/chief physician *Caroline Berg*
Journal author *Marianne Lindgren*
Discharge Date *20120325*
Care episode *20120311-20120318*
Main diagnosis acc to ICD-10 *DV073*
Medical history *52-year-old woman, familiar at the clinic. Goes to Karin* Lundgren and to the pain clinic.

Epikris *Huddinge sjukhus*
Ansv specialist-/överläkare *Caroline Berg*
Journalförare *Marianne Lindgren*
Utskriftsdatum *20120325*
Vårdtid *20120311-20120318*
Huvuddiagnos enl. ICD-10 *DV073*
Anamnes *52-årig kvinna, välkänd på kliniken. Går hos Karin Lundgren samt på smärtmottagningen.*

Figure 1: Example of a pseudonymised record. The original Swedish pseudonymised record is to the right and the translated version is to the left. The underlined words are the surrogates, where real data has been replaced with pseudonyms.

uses lexical, orthographic, syntactic and dictionary features. The CRF is based on trial-and-error experiments with feature sets described by Berg and Dalianis (2019), and uses the same features except for section features.

3.2.2 LSTM

The long short-term memory (LSTM) needs word embeddings as features for the training. Word2vec[5] was used to produce word embeddings using shallow neural networks, based on two corpora; a clinical corpus and medical journals. For the training using real clinical data, word embeddings were produced using a clinical corpus of 200 million tokens that produced 300,824 vectors with a dimension of 300.

For the training with pseudo clinical data, word embeddings were produced using Läkartidningen corpus (The Swedish scientific medical journals from 1996 to 2005) containing 21 million tokens that produced 118,662 vectors with a dimension of 300. The reason for using Läkartidningen is that the corpus does not contain sensitive data and hence is also more easily usable for transferable cross hospital training.

4 Results

The results of the experimental work are summarised in Figure 2. As can be observed in the figure, the CRF algorithm seems to generally outperform the LSTM algorithm on all metrics; precision, recall and F1 measure.

This result is not consistent with repeated reports in the literature, where deep learning approaches such as LSTM have been shown to out-perform most other methods, including CRF. Since deep learning approaches normally require very large amounts of data, one explanation for this result could be that the word embeddings used in this study did not contain sufficient context variations required for more robust performance or an insufficient training set of annotated data.

The ability to identify date part and age entities are similar when training on pseudonymised data and real data for the CRF. In contrast, *Location*, *Health Care Unit* and *Full Date* were negatively affected when using pseudonymised training data regardless of using a CRF or LSTM model.

4.1 CRF - Results

Experimental results of the CRF algorithm are shown in Table 3. Not presented in the table is the combination of training on real data and evaluation of pseudo data (*Real-Pseudo*), but the results of this combination gave a precision of 86.37 and recall of 77.80% and an F1-score of 81.86.

4.2 LSTM - Results

The experimental results of the LSTM algorithm are shown in Table 4 and again, not presented in the table is the combination of training on real data and evaluation of pseudo data (Real-Pseudo). The result of this combination is a precision of 65.83% and recall of 74.79% and F1-score of 70.03.

5 Analysis

The training set used in this study has a substantially constrained vocabulary compared to the evaluation set, which may partially explain the overall performance achieved when evaluating on real data (Pseudo-Real). The pseudo (training)

sklearn-crfsuite.readthedocs.io/en/latest/

[5] word2vec, https://github.com/tmikolov/word2vec

CRF	Real-Real			Pseudo-Pseudo			Pseudo-Real		
	P %	R %	F_1-score	P %	R %	F_1-score	P %	R %	F_1-score
First Name	95.94	92.42	94.15	98.52	98.08	98.30	97.14	72.39	82.96
Last Name	97.91	93.22	95.51	98.54	97.55	98.04	96.80	38.90	55.50
Age	97.06	68.75	80.49	100.00	68.09	81.01	97.50	81.25	88.64
Phone Number	94.69	82.95	88.43	92.37	80.15	85.83	83.48	74.42	78.69
Location	80.85	58.46	67.86	93.27	74.05	82.55	57.38	53.85	55.56
Full Date	95.68	95.48	95.58	91.02	86.32	88.61	47.56	21.97	30.06
Date Part	96.27	94.94	95.60	98.29	96.05	97.16	87.04	94.79	90.75
Health Care Unit	85.40	64.00	73.17	93.75	87.50	90.52	45.29	16.30	23.97
Overall	94.03	85.30	89.45	96.31	92.22	94.22	80.44	49.83	61.54

Table 3: Entity-based evaluation for CRF with ten fold cross-validation. A comparison is made for the different combinations of training on real data and evaluation on real (Real-Real) as well as pseudo data and on training on pseudo data and evaluation on pseudo (Pseudo-Pseudo) as well as real data (Pseudo-Real).

LSTM	Real-Real			Pseudo-Pseudo			Pseudo-Real		
	P %	R %	F_1-score	P %	R %	F_1-score	P %	R %	F_1-score
First Name	91.61	86.49	88.98	81.41	78.27	79.81	73.42	72.99	73.20
Last Name	96.40	87.02	91.47	89.29	91.88	90.57	84.70	75.00	79.55
Age	87.50	58.33	70.00	80.95	36.17	50.00	83.33	31.25	45.45
Phone Number	33.53	82.22	47.64	64.83	71.21	67.87	30.87	71.32	43.09
Location	20.47	46.02	28.34	60.71	17.35	26.98	27.40	10.77	15.47
Full Date	77.67	74.06	75.82	67.76	72.20	69.91	52.58	23.00	32.00
Date Part	90.31	90.60	90.45	91.48	95.08	93.25	59.08	94.79	72.79
Health Care Unit	68.37	61.82	64.93	81.24	81.63	81.44	27.18	14.00	18.48
Overall	76.76	78.62	77.68	82.49	81.79	82.14	60.56	55.10	57.70

Table 4: Entity-based evaluation for LSTM with three fold cross-validation where 66% of the data were used for training and 33% for evaluation. 10% of the data was previously held out as a development set. A comparison is made for the different combinations of training on real data and evaluation on real as well as pseudo data and on training on pseudo data and evaluation on pseudo as well as real data.

version of the data has less PHI tokens and the entities are more often single tokens.

The Full Date structure *yyyyddmm - yyyyddmm* is commonly occurring in the pseudo data, and the dash between the dates, "-", is often incorrectly identified. For example, using the CRF algorithm on real-data training and pseudo-data testing (Real-Pseudo), of the 159 instances not identified as full dates tokens, sixty contain '-'. The pseudo data uses the structure *yyyyddmm* while the real data uses *yyddmm*, which leads to errors. For these kinds of errors on standard data formats such as dates, it is easy to see how rule-based approaches using regular expressions could significantly improve the overall performance of the system.

The weakest performance area was for location information. There is a large variety of locations in the pseudo-data. These are also fairly specific and unlikely to occur in the real data, for example, locations with very few inhabitants. These uncommon rural places have names similar to residential homes *(äldreboenden)*. There are multiple instances of the suffix *'gården'* (yard) in the location pseudo-PHI, whereas, in the real data, the same suffix is common for care units.

In the pseudo-data, the care units are more general than in the real data, often too general to be annotated in the real data set. Infirmaries are fairly common in the real data but non-existent in the pseudo data. This lack of variation in the pseudo is partially responsible for the drop in performance.

There are at least two ways to think about mitigating this poor performance. First, location and care unit could be combined as one entity type since they are conceptually very similar, and sometimes have interchangeable entity names. Secondly, using more detailed municipality street and location mapping databases as dictionaries could be considered.

6 Discussion

There is one similar study to ours but for English by Yeniterzi et al. (2010), where the authors train their de-identification system with all combinations of pseudonymised textual data (or what they call resynthesized records) and real data and their results are in line with ours. However, there are some studies on cross-domain adaptation. In cross-domain adaption there is, however,

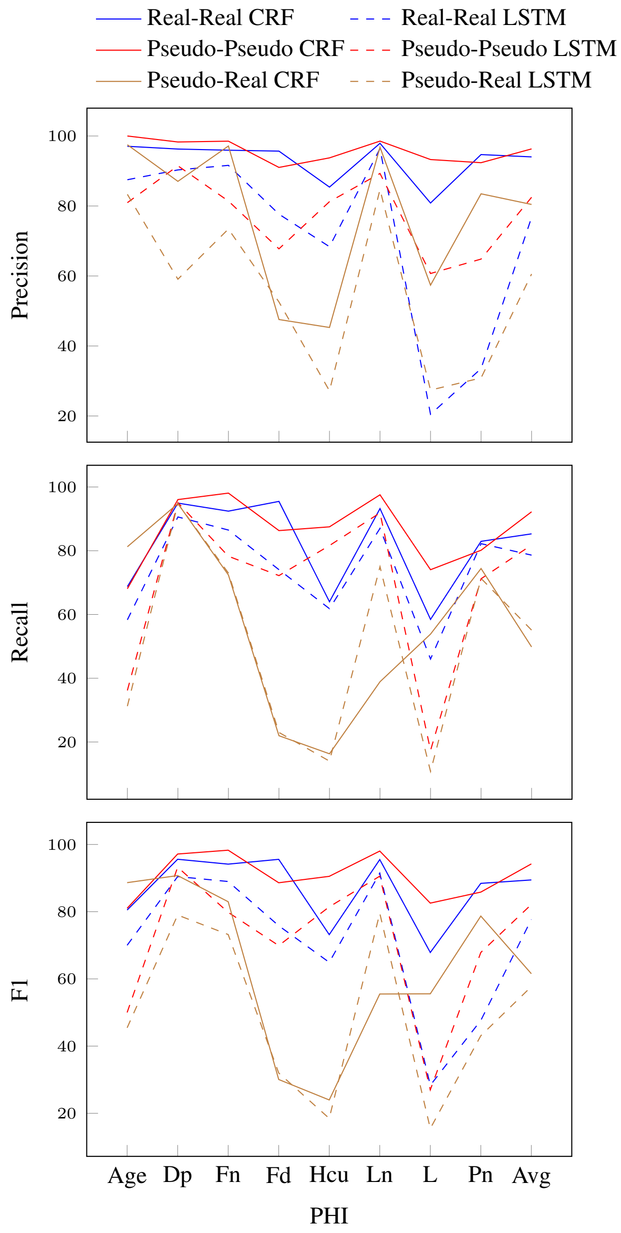


Figure 2: Line graphs visualising the results of both CRF and LSTM, and the outcomes of the evaluations. The x-axis have the PHI entities; Age, Date Part (Dp), First Name (Fn), Full Date (Fd), Health Care Unit (Hcu), Last Name (Ln), Location (L) and the average result (Avg).

a substantial domain change between the training and testing data, unlike in this study. Martinez et al. (2014) used models trained in one hospital on pathology reports in another hospital. Their system only required minor feature normalisation, and the reported results were comparable across the hospitals. Although this demonstrates feasibility, it is important to note that the pathology reports were from the same medical sub-speciality with only some narrative differences.

In this study, in addition to narrative differences between the training data and the target evaluation data, the number of care units and locations involved, as well as personal names, are widely varied. With large amounts of out of vocabulary variation, training on limited data will likely yield poor results. In practice, these data types exist in other non-sensitive sources such as city and rural location and street mapping data.

Except for location and care unit, evaluation on pseudo-data (Pseudo-Pseudo) produced better outcomes compared to performance on real-data (Pseudo-Real), which can be expected. What was a bit unexpected was the lower performance of the LSTM algorithm. The algorithm's results would potentially have been improved by larger vector data or more labelled data (Dernoncourt et al., 2017). While clinical notes have unique linguistic structures and grammatical peculiarities, non-clinical data sources could still provide important contextual information for constructing a useful vector space. Additional sources using non-sensitive data, such as public corpora in the general domain, hold a potential to improve performance on the de-identification task, therefore this line of inquiry will be followed up on in future work. In the same vein, factoring in part of speech tags from other sources of clinical data could be useful in this case. For instance, there are de-identification databases of clinical text, such as MIMIC (Neamatullah et al., 2008a; Goldberger et al., 2000), which could be used as additional information for training purposes, and using only the part of speech tags reduces security risks (Boström and Dalianis, 2012).

Current results are calculated as exact matches, and the partial match is not factored in, which may affect the result. As mentioned in the analysis the CRF algorithm rarely classifies the '-' in between dates as a part of the dates, and these are therefore not counted as matches despite the most identifying parts of the entity being identified.

To improve the general performance, a combination of both the LSTM and CRF algorithms could be performed instead of testing them independently. Combining high-performance algorithms and the use of ensemble methods seem to produce the best results as reported in the literature (Dernoncourt et al., 2017; Liu et al., 2017), and these techniques will be investigated in future work on the data sets.

7 Conclusions and future directions

The results of this study suggest that although it is possible to train models on pseudonymised data for use in different contexts, there is severe deterioration in performance for some PHI information. Even small narrative and distributional variation could negatively impact performance.

Transferring a system from one set of clinical text to a different set could result in the performance of the system deteriorating; in this study the Pseudo-Real case. This problem, what we call *The cross pseudo-real text adaptation problem*, is an issue that could happen due to the pseudonymisation/de-identification processes on the training data due to the narrative and distributional variation as well as other differences in the nature of the PHI between the training data and the target.

In the future, we will try to improve the pseudonymisation module described in Dalianis (2019) to produce a larger variation in the vocabulary as the lack of variation may affect the current result negatively.

We will also apply the learned models to other Nordic languages such as Norwegian clinical text and use the system as a pre-annotation system to assist the manual annotators in their work to create a Norwegian gold standard.

Acknowledgments

We are grateful to the DataLEASH project and Helse Nord for funding this research work.

References

Hanna Berg and Hercules Dalianis. 2019. Augmenting a De-identification System for Swedish Clinical Text Using Open Resources (and Deep learning). In *Proceedings of the Workshop on NLP and Pseudonymisation, NoDaLiDa, Turku, Finland September 30, 2019*.

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.

Henrik Boström and Hercules Dalianis. 2012. De-identifying health records by means of active learning. In *Proceedings of ICML 2012, The 29th International Conference on Machine Learning*, pages 1–3.

Hercules Dalianis. 2019. Pseudonymisation of Swedish Electronic Patient Records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation, NoDaLiDa, Turku, Finland September 30, 2019*.

Hercules Dalianis and Henrik Boström. 2012. Releasing a Swedish clinical corpus after removing all words–de-identification experiments with conditional random fields and random forests. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC*, pages 45–48.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. HEALTH BANK–A Workbench for Data Science Applications in Healthcare. In *Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015), J. Krogstie, G. Juel-Skielse and V. Kabilan, (Eds.), Stockholm, Sweden, June 11, 2015, CEUR, Vol-1381*, pages 1–18.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying Swedish clinical text - Refinement of a Gold Standard and Experiments with Conditional Random fields. *Journal of Biomedical Semantics*, 1:6.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Ary L. Goldberger, Luciani Alano Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chih Kang Peng, and Harry Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101 23:E215–20.

Clete A. Kushida, Deborah A. Nichols, Rik Jadrnicek, Ric Miller, James K. Walsh, and Kara Griffin. 2012. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical Care*, 50(7):S82–S101.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, 75:S34–S42.

David Martinez, Graham Pitson, Andrew MacKinlay, and Lawrence Cavedon. 2014. Cross-hospital portability of information extraction of cancer staging information. *Artificial intelligence in medicine*, 62(1):11–21.

Andrew J. McMurry, Britt Fitch, Guergana Savova, Isaac S. Kohane, and Ben Y. Reis. 2013. Improved de-identification of physician notes through integrative modeling of both public and private medical text. *BMC medical informatics and decision making*, 13:112–112. 24083569[pmid].

Stephane Meystre, Jeffrey Friedlin, Brett South, Shuying Shen, and Matthew Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10(1):70.

Ishna Neamatullah, Margaret M. Douglass, Li-wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. 2008a. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. 2008b. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):1.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields. Accessed 2019-06-17.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of biomedical informatics*, 75:S4–S18.

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. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Reyyan Yeniterzi, John Aberdeen, Samuel Bayer, Ben Wellner, Lynette Hirschman, and Bradley Malin. 2010. Effects of personal identifier resynthesis on clinical text de-identification. *Journal of the American Medical Informatics Association*, 17(2):159–168.

Vithya Yogarajan, Michael Mayo, and Bernhard Pfahringer. 2018. A survey of automatic de-identification of longitudinal clinical narratives. *arXiv preprint arXiv:1810.06765*.