Augmenting a De-identification System for Swedish Clinical Text Using Open Resources and Deep Learning

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

Hercules Dalianis
Department of Computer and Systems Sciences
Stockholm University
hercules@dsv.su.se

Abstract

Electronic patient records are produced in abundance every day and there is a demand to use them for research or management purposes. The records, however, contain information in the free text that can identify the patient and therefore tools are needed to identify this sensitive information.

The aim is to compare two machine learning algorithms, Long Short-Term Memory (LSTM) and Conditional Random Fields (CRF) applied to a Swedish clinical data set annotated for de-identification. The results show that CRF performs better than deep learning with LSTM, with CRF giving the best results with an F_1 score of 0.91 when adding more data from within the same domain. Adding general open data did, on the other hand, not improve the results.

1 Introduction

Electronic health records (EHR) are today produced in abundance and consist of information valuable to improve the medical care of future patients. They are, however, seldom reused for research as free text in patient records often contain possibly identifiable information about patients. To enable access to electronic health records while preserving patient privacy there is a need for automatic de-identification.

The US Health Insurance Portability and Accountability Act (HIPAA) defines 18 categories of Protected Health Information (PHI) which has to be concealed for EHRs to be considered de-identified in the US (Health Insurance Portability and Accountability Act (HIPAA), 2003). The categories include names, geographic divisions smaller than state, dates related to an individual, contact information and other data that can uniquely identify the individual.

Modules built to identify PHI, primarily rely on two methods: Rule-based methods and supervised machine-learning methods (Meystre et al., 2010). The two methods are often used together in hybrid systems (Stubbs et al., 2017). Rule-based methods do not require annotated data for training, are easy to modify and the results are easy to interpret, but they lack robustness and designing rules is a complex task (Meystre et al., 2010). Machine learning methods may provide greater robustness, but require an abundant amount of annotated data. According to Dernoncourt et al. (2017), statistical machine learning models require feature engineering, while artificial neural networks (ANN) does not. The latter does, however, require more data.

Lee et al. (2017) show that training a model on a large source dataset and then fine-tuning by retraining it on the smaller target data set can improve the results in comparison to only using the smallest data set. While the data sets used by Lee et al. (2017) consisted of 29,000 PHI instances in the smaller target data set and 61,000 PHI instances in the larger source data set the largest available Swedish data set, the Stockholm EPR PHI Corpus, has only 4,421 instances of PHI (Velupillai et al., 2009; Dalianis and Velupillai, 2010). It does exist a smaller related corpus with Electronic Health Records with annotations for de-identification, the Stockholm EPR PHI Domain Corpus (Henriksson et al., 2017b). For a larger data set with general Swedish text annotated for named entity recognition, Stockholm Umeå Corpus exists (Östling, 2012).

This study investigates the possibilities of augmenting the quality of de-identification by adding a general Swedish data set for named entity recognition such as Stockholm Umeå Corpus to already existing annotated PHI data sets and secondly the

Proceedings of the Workshop on NLP and Pseudonymisation, pages 8–15,
Turku, Finland, 30 September 2019.

use of deep learning methods such as LSTM.

2 Previous research

The state-of-the-art de-identification systems have for a long time been hybrid systems, where a machine learning approach, typically Conditional Random Fields (CRF) is used to identify classes including names, professions, and locations and a rule-based approach is used to identify rarely occurring or regular classes as zip codes, phone numbers and e-mail addresses (Uzuner et al., 2007; Stubbs et al., 2015). The best result during the i2b2 de-identification challenge 2014 (Stubbs et al., 2015) has a micro-averaged entity-based recall of 93.90%, a precision of 97.63% and an F_1 score of 0.96 on i2b2 PHI-categories.

The first neural network de-identification system was introduced in 2016 (Dernoncourt et al., 2017). This system used a type of deep learning with recurrent neural networks (RNN) called long short-term memory (LSTM) with three layers: A character enhanced token-embedding layer, a label prediction layer and a label sequence optimisation layer. The model is bidirectional to better handle long term dependencies. The ANN model presented, performed better than the best system from the i2b2 2014 challenge. Combining Bi-LSTM and CRF further improved the system. Similar systems based on LSTM and CRF have been successful for de-identification (Liu et al., 2017), and during the i2b2 de-identification challenge of 2016 a model combining an LSTM, a CRF and rules won the challenge with an entity-based micro-averaged F_1 score of 0.91 for HIPAA classes (Stubbs et al., 2017).

The largest Swedish dataset with health records annotated for de-identification is the Stockholm EPR PHI Corpus, which is a part of Health Bank - Swedish Health Record Research 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 first results for identifying PHI based on the gold standard of the Stockholm EPR PHI Corpus can be seen in Table 1. De-identification tasks based on CRF as well as rules have been carried out on this data set with precision scores between 85% and 92.65%, recall scores between 71% and 81% and F_1 scores between 0.76 and 0.87 (Dalianis and Velupillai, 2010; Henriksson et al., 2017b; Dalianis and Boström, 2012; Boström and Dalianis, 2012). The best de-identification system based on the corpus was developed by Henriksson et al. (2017b), using token, lemma, part of speech, capitalisation, digit, compounds, and dictionary matches against the medical terminologies SNOMED CT, MeSH as features. Predictive performance estimates yielded an F_1 score of 0.87.

McMurry et al. (2013) have trained decision tree classifiers using 28 features based on part of speech tags, term frequencies, and dictionaries in open journal publications and confidential physician notes to recognise non-PHI words. According to the study, distributional differences between private and open medical texts can be used to classify PHI.

3 Data and method

3.1 Data

Three data sets for de-identification are used: The Stockholm EPR PHI Corpus, the Stockholm EPR PHI Domain Corpus and Stockholm Umeå Corpus 3.0 (SUC). The data consists of both clinical data[1] and open-source data. The Stockholm EPR PHI Corpus is used both for development, training, and testing, while Stockholm EPR PHI Domain Corpus and SUC are only used for training.

All data is encoded using BIOES-encoding, indicating the position of the token within the PHI entity. It encoded whether the token was in the **B**eginning, **I**nside or **E**nding of a multi-token entity, a **S**ingle entity or **O**utside an entity (Reimers and Gurevych, 2017).

Stockholm EPR PHI Corpus consists of 100 patient records from five clinical units: Neurology, orthopaedia, infection, dental surgery and nutrition at Karolinska University Hospital (Dalianis and Velupillai, 2010) and has approximately 200,000 tokens. The Stockholm EPR PHI Corpus was first manually annotated by three annotators into 28 PHI classes based on HIPAA and enriched with further classes (Velupillai et al., 2009). The annotations were later on merged into conceptually similar classes while removing classes with few instances, creating a gold standard with eight PHI annotation classes: *Age, numeric and non-numeric full dates and*

[1] This research has been approved by the Regional Ethical Review Board in Stockholm (2012/834-31/5).

				Exact matches			Partial matches		
Class	**Annotated**	**Retrieved**	**Relevant**	**Precision**	**Recall**	**F-score**	**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)

date parts, first names, last names, health care units, locations, and *phone numbers* (Dalianis and Velupillai, 2010). Locations include not only places but also companies. Health care units were only annotated as Health Care Unit if they were considered identifiable by the annotator. The distribution of PHI is presented in Table 2.

Stockholm EPR PHI Domain Corpus consists of data from three clinical units: Geriatric, oncology and orthopaedic at Karolinska University Hospital. It has approximately 116,000 tokens. It uses the same eight annotation classes as the Stockholm EPR PHI Corpus. In the original version, almost half of the corpus is annotated, while the other half is not. The original annotation for health care unit followed other guidelines than the one set in (Dalianis and Velupillai, 2010). The Health Care Unit annotations and other half of the corpus were therefore re-annotated in this study. Health care units were only annotated if they were identifiable within the Stockholm area.

Stockholm Umeå Corpus 3.0 consists of Swedish texts from press, scientific writing and prose collected during the 1990s and has over one million tokens (Östling, 2012; Gustafson-Capková and Hartmann, 2006). The latest release was SUC 3.0, released in 2012. The corpus is annotated with part-of-speech tags, morphological analysis, lemma as well as ten named-entity classes. The used classes are *person, place, institution, animal, myth*[2], *product, work, measurements (with age as a subclass), event* and *other*. The annotations for person, location and age were used in this study, further the person annotation was semi-manually divided into first names and last names. The entire corpus is used.

	EPR	**Domain**	**SUC**
First Name	928	380	11,748
Last Name	923	524	9,402
Phone Number	135	47	0
Age	56	52	427
Full Date	500	382	0
Date Part	710	555	0
Health Care Unit	1,021	387	24
Location	148	96	9,388
Total	4,421	2,886	30,989

Table 2: Overview of annotated Protected Health Information entities. Note that Date Parts, Full Dates or Phone Numbers are not annotated in SUC.

The Stockholm EPR PHI Corpus was first divided into two sets: One small for development and validation with 10% of the patient records and one for training and testing by cross-validation with 90% of the patient records. For the CRF, tenfold cross-validation was used. The patient records from the Stockholm EPR PHI Corpus were divided into ten folds. The Stockholm EPR Domain Corpus and SUC Corpus were divided into ten folds, where for each fold 90% of the sentences were used for training. Only the folds from the Stockholm EPR PHI Corpus were used for testing. A similar approach was done for LSTM,

[2]The myth annotation consists of names of mythical creates and places and the animal annotation consists of names of animals.

but used validation data for early stopping. The LSTM has only been evaluated on the three first folds due to time constraints.

3.2 Method

This study compares the predictive powers for three models based on the data described above. The first model is only trained on data from the Stockholm EPR PHI Corpus, the second model is trained on data from the Stockholm EPR PHI Corpus and the Stockholm EPR Domain Corpus. The last model is trained on the Stockholm EPR PHI Corpus and SUC. All models are evaluated on data from the Stockholm EPR PHI Corpus using ten cross fold-validation.

The result is evaluated with micro averaged entity-based precision, recall and F_1 score, which is the standard for evaluating named entity recognition (Stubbs et al., 2015).

3.2.1 LSTM

Recurrent neural network (RRN) is a type of deep learning artificial neural network designed for processing sequential data (Dernoncourt et al., 2017). The bidirectional LSTM architecture is designed to access long-range dependencies in both forward and backward directions (Dernoncourt et al., 2017). The experiment uses the architecture described in Lample et al. (2016) based on an open-source implementation with Tensorflow[3].

As stated by Lample et al. (2016), character-based representations can be used to capture both morphological and orthographic information. The character-representations are learned from the used training set for each experiment. Pre-trained word representation is used, based on a subset of clinical text from Health Bank of 200 millions tokens producing 300,824 vectors with a dimension of 300.

The implementation uses the adaptive learning rate method *Adam*, an algorithm for optimisation of stochastic objective functions (Kingma and Ba, 2014). It computes different learning rates for each parameter based on estimates from the first and second moments of the gradients. The *learning rate* was set to 0.001 with a *decay of 0.9*.

Dropout was used with a *dropout rate of 0.5*. This was used with a *batch size of 64*. The training is done in a maximum of *20 epochs*, with early stopping if no improvement three times in a row in the development set. The model was then evaluated on the test set. The CRF layer (Lample et al., 2016) was not used as it did not show any benefits for the validation set compared to using only LSTM.

3.2.2 CRF

Conditional Random Fields (CRFs) with linear chain is a statistical machine learning method first introduced by Lafferty et al. (2001) that predicts sequences of labels based on sequences in the input. A set of features is typically defined to extract features for each word in a sentence. The CRF tries to determine weights that will maximise the likelihood of leading to the labels in the training data.

In this study, CRFSuite (Okazaki, 2007) is used with a the sklearn-crfsuite wrapper[4]. The features used are: Word as lower case, the first and last four and eight letters, lemma, part of speech tag, if the word is in lower case, upper case or title case, if there are only numbers in the word or only letters in the word or if it has special characters, how many letters, numbers or other characters the word has. This is carried out with a window size of 5. Information about which heading a word comes from is included for texts from the Stockholm EPR PHI Corpus. Furthermore, the CRF uses gazetteers for first names, last names, locations, honorifics or medical profession titles, hospitals in the Stockholm region and regular expressions for identifying date parts, full dates and telephone numbers.

The CRF uses gradient descent with Limited-memory BFGS (L-BFGS) for optimization. LBFGS is an optimization algorithm (Koller et al., 2007).

Lemma and part of speech tagging for each word was performed with Stagger (Östling, 2012) for the Stockholm EPR PHI Corpus and the Stockholm EPR Domain Corpus. SUC is already manually annotated with lemma and part of speech.

4 Results

4.1 LSTM - Results

As seen in Table 3 presenting the results for the LSTM, the systems handle first names, last names, date parts, full dates better than ages, health care units and location.

[3]https://github.com/guillaumegenthial/sequence_tagging

[4]https://sklearn-crfsuite.readthedocs.io

	EPR PHI			EPR PHI + Domain			EPR PHI + SUC		
	P %	R %	F_1	P %	R %	F_1	P %	R %	F_1
First Name	93.79	92.63	0.93	95.16	93.99	**0.95**	92.46	91.46	0.92
Last Name	93.09	94.87	0.94	97.19	95.11	**0.96**	90.77	96.68	0.94
Phone Number	96.30	94.44	**0.95**	90.00	95.83	0.93	95.24	91.67	0.93
Age	80.56	75.56	0.78	70.00	75.56	0.72	91.67	75.56	**0.82**
Full Date	91.38	96.46	0.94	91.98	95.77	0.94	92.59	95.82	0.94
Date Part	95.6	97.96	**0.97**	93.51	97.24	0.95	93.37	94.60	0.94
Health Care Unit	61.19	69.20	**0.65**	58.95	54.70	0.57	59.06	51.88	0.55
Location	76.90	75.27	**0.76**	69.87	70.15	0.69	61.55	86.54	0.70
Overall	85.87	88.96	**0.87**	86.28	85.46	0.86	84.78	85.44	0.85

Table 3: Entity-based evaluation for LSTM for the first three folds. The mean is presented for each label. The highest F_1 scores are highlighted for each class.

The only two types of PHI improved when adding SUC is *age* and *full date* and no improvements can be seen in any other classes. Rather a drop of performance can be seen for location, last names and first names. There is a small increase of recall for first names and locations, but with lower precision.

Stockholm EPR PHI Corpus alone performs considerably better for identifying phone numbers, locations and health care units, while first names and full dates seem to be identified correctly to a greater extent with additional data from the Stockholm EPR PHI Domain Corpus.

Overall, there is no improvement when adding another corpus to the training, but rather a drop in performance.

4.2 CRF - Results

Overall the CRF systems perform well, particularly for finding dates and names. The recall is lower for *Phone Number* and both the precision and recall is lower for *Health Care Unit, Location* and *Age*. As seen in Table 4 with results for the CRF, compared to LSTM results in Table 3, the CRF performs better overall with greater precision, but the LSTM has a higher recall.

Adding the Domain Data increases the F_1 score marginally. Some small, likely insignificant, improvements can be seen for *Health Care Unit*, *Age* and *Phone Number*. There is not the same drop of performance as for the LSTM systems.

Adding SUC does not improve the ability to predict, and instead both precision and recall is lower for all classes except last names and ages. The drop of performance is however less severe than for the LSTM.

5 Analysis

Health Care Unit and *Location* are the most commingled PHI classes. Health care units are often named by their geographic location. *Huddinge* can for example refer to the hospital *Karolinska University Hospital Huddinge* but also the municipality *Huddinge*. In the gold standard, locations are annotated as a part of the health care unit occasionally depending if it is an actual part of the name and whether it is directly adjacent to an identifiable health care unit. Errors are partly caused by the difficulty to distinguish these cases. Furthermore, some health care units are only occasionally annotated as PHI, which also makes it more difficult for the system to learn the structure. *ASIH*, which stands for Advanced Care At Home in Swedish, is for example in 8 of 20 cases annotated as a singular health care unit entity.

Location is a class with generally low F_1 score. One reason for this may be that the test data includes companies as locations. Location has relatively few annotations, and almost one-quarter of these are company annotations. Companies are overall rarely occurring, but frequently mentioned in one patient record. In the record with the most company annotations, none of the seven mentioned companies is found by any system.

When identifying age, the numeral in the age entity is often correctly identified, but the upcoming word is either incorrectly included or missed. The unit following the numeral, often 'years', is occasionally annotated within the PHI and occasionally not, which is one reason for these errors. Age annotations where the numeral is followed by 'årig' (year-old) are found to a greater extent than those followed by 'år' (years).

	EPR PHI			EPR PHI + Domain			EPR PHI + SUC		
	P %	R %	F_1	P %	R %	F_1	P %	R %	F_1
First Name	95.05	92.78	**0.94**	95.50	91.41	0.93	94.51	92.24	0.93
Last Name	97.02	92.20	0.94	96.39	90.36	0.93	96.93	0.93	**0.95**
Phone Number	92.81	81.52	0.87	94.58	84.32	**0.89**	96.14	72.14	0.82
Age	79.29	60.95	0.68	85.09	71.27	0.77	89.67	76.21	**0.82**
Full Date	98.62	99.15	**0.99**	96.34	94.74	0.96	95.82	93.28	0.94
Date Part	97.06	95.68	**0.96**	97.45	94.73	**0.96**	95.46	90.08	0.93
Health Care Unit	86.11	66.40	0.75	88.62	72.79	**0.80**	85.45	67.38	0.75
Location	74.07	73.70	**0.72**	76.05	59.89	0.66	62.40	70.92	0.65
Overall	93.76	86.53	0.90	94.66	86.72	**0.91**	92.31	84.97	0.88

Table 4: Entity-based evaluation for CRF with tenfold cross-validation. The mean is presented for each label. The highest F_1 scores are highlighted for each class.

Uncommon names, common words that are also names, misspelt names and names in lower case are less often identified. This especially happens in contexts where there are no other words, either to the left or right. This is common in sections similar to 'Assigned nurse'. There are some cases where first names are annotated as last names and vice verse. The first names *Carina, Riita* and *Abdul* are annotated as last names in the gold standard, leading to errors.

Non-PHI adjacent to a PHI entity is annotated as PHI and more general entities, similar to annotated PHI, for example "the summer of 2007", are more often mistaken as PHI. There are also some cases where inconsistent annotation leads to false positives for especially Health Care Units, they also lead to false negatives. The systems also manage to find some PHI not previously annotated in the gold standard.

6 Discussion

In comparison to other work where the Stockholm EPR PHI Corpus is used to classify PHI this set of features and CRF implementation works well for identifying PHI. The CRF also performed better by itself than the LSTM when focusing on the overall F_1 score. This the highest recall overall of 88.96% is nonetheless achieved with the LSTM system and without any additional corpus.

There is an overall drop of recall and F_1 when adding other corpora to the LSTM version, while the CRF version is slightly improved by adding the EPR PHI Domain corpus, with an F_1 score of 0.91. On the other hand the highest recall for *Last name* is achieved using LSTM and SUC and the recall of *First name* and *Last Name* is also improved with additional EPR data in. It could be argued that recall is more important than precision and and that *Last name* and *First name* are two of the most sensitive classes.

In SUC, organisations and places are annotated separately. Company names tend to sound like names and occur in similar contexts like health care units. A distinction between locations and companies may enable the usage of the organisation annotation from SUC, with possible improvements on similar labels as well as reduce the heterogeneity.

Differences between the annotation quality or guidelines may also affect the result. Inaccuracies within the Stockholm EPR PHI Corpus is mentioned in the analysis. The Stockholm EPR PHI Corpus was annotated by three annotators and further examined by others. The Stockholm EPR Domain Corpus was, however, originally annotated by only one person and re-annotated for this study by one of the authors to comply with the annotation guidelines of the Stockholm EPR PHI Corpus. This corpus is likely to have more inaccuracies than the Stockholm EPR PHI Corpus.

There is generally a drop in performance between domains and within cross clinical or cross hospital settings. Therefore, it may not come as a surprise that training partially on another domain does not benefit the classifier regardless of the data size. Open text within the medical domain may be more beneficial due to higher domain similarities. A selection of specific documents within SUC is unlikely to benefit the classifier as only a minority of SUC includes medical text.

Using partial match may improve the results for multi-token entity expressions, such as phone

numbers, locations, dates and health care units, see Figure 1.

7 Conclusion and future directions

This study aimed to investigate the possibilities of augmenting the quality of de-identification by using annotated data sets for named entities or the use of deep learning methods such as LSTM. The findings suggest that adding data from a general corpus for named entities is not a viable option, but perhaps for individual classes. LSTM performs reasonably well by itself, even if the CRF models seem to perform better. It is worth noting that the LSTM is not yet evaluated on all folds, and considering the increase of recall, it is still warranted to see if a hybrid version of this CRF and LSTM can improve the results further. One possible approach would be to use a LSTM system to de-identify personal names and a CRF system to de-identify phone numbers, locations, dates and health care units.

The current study only examined the effects of using two corpora together as training data, and not the performance when training on one data set, the Stockholm EPR PHI Corpus or SUC, and then using domain adaptation to the target date set, the Stockholm EPR PHI Corpus. While the identification of some PHI classes benefit from added data, there are also classes where no improvements are seen despite data being added.

The analysis has shown that there is a need to revise the old gold standard for the Stockholm EPR PHI by adding previously overlooked PHI, changing PHI accidentally annotated as another PHI, and possibly review the guidelines for the manual annotation of health care units, locations and ages.

Our best performing de-identification system surpasses previous systems based on Stockholm EPR PHI Corpus. It performs in line with the best performing de-identification systems from the latest i2b2 de-identification challenge (Stubbs et al., 2017) but lower than the best from earlier challenge (Stubbs et al., 2015). One observation, however, is that data set in Stubbs et al. (2015) is seven times larger than the Stockholm EPR PHI Corpus in terms of both tokens and PHI instances (Dernoncourt et al., 2017).

Acknowledgments

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

References

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 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.

Sofia Gustafson-Capková and Britt Hartmann. 2006. Manual of the Stockholm Umeå Corpus version 2.0. *https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf*.

Health Insurance Portability and Accountability Act (HIPAA). 2003. http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm U.S. Department of Health and Human Services. Accessed 2019-06-17.

Aron Henriksson, Maria Kvist, and Hercules Dalianis. 2017b. Detecting Protected Health Information in Heterogeneous Clinical Notes. volume 245, pages 393–397.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Daphne Koller, Nir Friedman, Sašo Džeroski, Charles Sutton, Andrew McCallum, Avi Pfeffer, Pieter Abbeel, Ming-Fai Wong, David Heckerman, Chris Meek, et al. 2007. *Introduction to statistical relational learning*. MIT press.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann.

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

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Transfer learning for named-entity recognition with neural networks. *arXiv preprint arXiv:1705.06273*.

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.

Andrew J. McMurry, Britt Fitch, Guergana Savova, Isaac S. Kohane, and Ben Y. Reis. 2013. https://doi.org/10.1186/1472-6947-13-112 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.

Naoaki Okazaki. 2007. http://www.chokkan.org/software/ crfsuite CRFsuite: a fast implementation of Conditional Random Fields. Accessed 2019-06-17.

Robert Östling. 2012. Stagger: A modern POS tagger for Swedish. In *The Fourth Swedish Language Technology Conference, Lund, Sweden*.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

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.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 78(12):e19–e26.