Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4444–4450
Marseille, 11–16 May 2020

A Semi-supervised Approach for De-identification of Swedish Clinical Text

Hanna Berg, Hercules Dalianis
Department of Computer and Systems Sciences
Stockholm University
{hanna.berg, hercules}@dsv.su.se

Abstract

An abundance of electronic health records (EHR) is produced every day within healthcare. The records possess valuable information for research and future improvement of healthcare. Multiple efforts have been done to protect the integrity of patients while making electronic health records usable for research by removing personally identifiable information in patient records. Supervised machine learning approaches for de-identification of EHRs need annotated data for training, annotations that are costly in time and human resources. The annotation costs for clinical text is even more costly as the process must be carried out in a protected environment with a limited number of annotators who must have signed confidentiality agreements. In this paper is therefore, a semi-supervised method proposed, for automatically creating high-quality training data. The study shows that the method can be used to improve recall from 84.75% to 89.20% without sacrificing precision to the same extent, dropping from 95.73% to 94.20%. The model's recall is arguably more important for de-identification than precision.

Keywords: semi-supervised learning, self training, de-identification, Swedish clinical text

1. Introduction

An abundance of electronic health records is produced every day. They possess valuable information for research and future improvement of medical care. A great majority will, however, never be re-used for research. This is partly due to the presence of sensitive data. Sensitive data may reveal the identity of individual patients. Sensitive data also restricts the possibility to re-use the valuable information in electronic health records for research. Sensitive data must, therefore, be removed to protect the integrity of patients.
Multiple efforts have been done to protect the integrity of patients while making electronic health records usable for research by removing personally identifiable information in patient records (Velupillai et al., 2009; Stubbs et al., 2017; Dernoncourt et al., 2017). Personally identifiable information may, for example, be personal names, addresses, phone numbers, dates and locations. Standard methods for de-identification is Named Entity Recognition by identifying personally identifiable information and then removing them (Meystre, 2015). The American HIPAA Privacy Rule states that a health record is de-identified if eighteen types of identifiers are removed from the record (HIPAA, 2003). De-identification systems are designed to identify and remove this personally identifiable information (Velupillai et al., 2009; Stubbs et al., 2015).

The annotation process for supervised machine learning methods is costly (Haertel et al., 2008). The annotation costs for clinical text is even more costly as the process must be carried out in a protected environment with a limited number of annotators who must have signed confidentiality agreements. Dernoncourt et al. (2017) estimated, from numbers presented in (Neamatullah et al., 2008)[1], that manual de-identification of the clinic MIMIC dataset, would require at least two annotators and a total of 5,000 hours to annotate the whole 100-million-word dataset. Semi-supervised learning would, on the other hand, enable the usage of unlabelled electronic health records and possibly reduce annotation costs while improving de-identification accuracy.

An approach for improving annotation speed is described in Hanauer et al. (2013). The study included bootstrapping and iterative pre-annotation to make annotation faster and the annotation speed was doubled with preserved quality. Lingren et al. (2013) similarly showed that pre-annotation may save time, but from 13.85% to 21.50% for each entity. Rehbein et al. (2009), however, found no conclusive evidence that pre-annotation can speed up the annotation process, but it can increase the quality.
Semi-supervised learning is a research area that covers several sub-areas, for example: Active learning, pre-annotation, bootstrapping, co-training and self-learning. The goal of semi-supervised learning is to take advantage of unlabelled data to extend the labelled data. The approach carried out in this article is semi-supervised learning with self-training with machine annotated entities, studying whether it can be an option to annotating more data.
In this paper, the research question addressed is: Is it possible to use semi-supervised learning to obtain more high-quality training data?

2. Previous research

Self-training is likely the easiest method of using unlabelled data and one of the first attempts of semi-supervised learning. Self-training essentially starts with building a single classifier with labelled data, and then iteratively label unlabelled data (Nigam and Ghani, 2000). The newly labelled predictions are combined with the actual training data, treating the predictions as the truth and used to label more unlabelled data (McClosky et al., 2006). Self-training is normally not very effective. Since the classifier each time uses its predictions to teach itself, there is a considerable

[1] Neamatullah et al. (2008) implemented a rule-based Deid-system in the programming language Perl. The system obtained a recall of 0.943.

risk that mistakes are reinforced throughout and amplified.

The *Yarowsky Algorithm* is an example of self-training, which was initially used for word sense disambiguation (Yarowsky, 1995). It makes the assumption that it is unlikely for multiple occurrences of the same word to have different meanings in the same discourse. This assumption is used to select words with high confidence and then adding words in the same discourse with the same sense to the training data despite lower confidence.

Self-training has also been shown to improve results when used for adapting data from one domain to another since it does not rely as much on existing annotations (McClosky et al., 2006). Wu et al. (2009) introduced a domain adaptive bootstrapping method with a selection criterion relying on finding domain-specific and domain-independent non-general instances which may work as a bridge between the domains.

There are cases where self-training may lead to a substantial deterioration of the accuracy in a system (Zhou et al., 2012). If the unlabelled data favours one particular class of data, the risk of over-fitting increases. Self-training also likely to introduce noise, and with too much noise the classifier's accuracy will deteriorate. There is also a risk that the selected unlabelled data will not add more information, leading to no improvements.

In theory, if the data size is enough the benefits of augmented labelled data will over-weigh the noise introduction (Zhou et al., 2012). It is, however, in practice close to impossible to estimate the required size as the exact hypothesis space will be unknown. For that reason, it is necessary to find the right data, which often is done by choosing data which likely is labelled correctly. An overly conservative model, however, risks not adding new knowledge. Rosenberg et al. (2005) have applied self-training to object detection systems with improvements over state-of-the-art systems and benefits has also been seen for named entity recognition (Kozareva et al., 2005).

Co-training is different from self-training in the sense that it uses multiple classifiers with distinctive, different feature sets with different views of the data. The predictions of each classifier are used as an additional input to the other classifier during each classifier. The intuition is that each classifier provides extra, useful information to each other while reducing the risk of amplified errors. According to Blum and Mitchell (1998) co-training assumes that the two views are individually sufficient for classification and that the two views are conditionally independent of each other. Later research argues that a weaker independence assumption still holds (Dasgupta et al., 2002). A variation of this is, for example, using different learning algorithms rather than the same algorithm with different feature sets.

A study with active learning on the Stockholm EPR PHI Corpus showed that selecting the most uncertain examples resulted in lower predictive performance than random selection or selecting the most certain examples (Boström and Dalianis, 2012). Selecting the most certain examples seemed to perform the best.

3. Methods and Data

3.1. Data

3.1.1. Stockholm EPR PHI Corpus

Stockholm EPR PHI Corpus[2] consists of 98 patient records in Swedish from five clinical units at Karolinska University Hospital: *Neurology, orthopaedia, infection, dental surgery* and *nutrition* containing approximately 200,000 tokens in total (Dalianis and Velupillai, 2010). The corpus was manually annotated by three annotators into 28 identifier classes and is intended for training Named Entity Recognition for de-identification systems (Velupillai et al., 2009). Some of the 28 classes contained very few instances and therefore were the 28 classes later merged into eighth conceptually similar classes to be of practically use. This process is described in (Dalianis and Velupillai, 2010). The eight classes used in this study are: *Age, Full date, Date part, First name, Last name, Health care unit, Location*, and *Phone number*. The data only includes free text. The distribution is imbalanced with roughly 3-4% of all tokens being part of a personally identifiable entity and only 25% of all sections includes at least one personally identifiable entity.

For this study the data from Stockholm EPR PHI Corpus was divided into three sets: Training set (60%), development set (10%) and test set (30%). The training set is hereby referred to as L and the test set as E. Both of these are manually annotated. The data in the training set and test set are from different clinical units.

The labelled training data consists of 95,500 tokens from the Stockholm EPR PHI Corpus. The data is from 2008 and includes patient records from 62 patients. The training data uses electronic records from all clinical units except the infection unit.

The labelled evaluation data consists of 54,700 tokens from the Stockholm EPR PHI Corpus. The data is from 2008 and includes records from 36 patients. The data comes from an infection clinic.

10% of the data is used within the bootstrapping cycle to determine the selection and stopping criterion. The data consists of 20,488 tokens from a nutrition clinical unit.

3.1.2. Health Bank

The research infrastructure Health Bank[3] (The Swedish Health Record Research Bank) where also Stockholm EPR PHI Corpus is contained, encompasses also a considerably larger corpus with records from 512 clinical units from Karolinska University Hospital from over two million patients (Dalianis et al., 2015). The whole corpus contains over 500 million tokens and includes both structured and unstructured information.

The unlabelled data used in this study is a subset of Health Bank of 100,000 sentences with a total of around 2,000,000 tokens. The data comes from a collection of different clinical units including cardiologist units, the language and speech pathology clinic and orthopaedic units and is mainly from 2008 and 2009. The unlabelled dataset contains 20 times more tokens than the training data. There may be an

[2]This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

[3]Health Bank, `http://www.dsv.su.se/healthbank`

overlap in clinical units between the labelled and unlabelled data with the same medical professionals occurring in both the unlabelled and labelled data.

3.2. Conditional Random Fields

Conditional Random Fields (CRFs) with linear chain is a probabilistic framework for labelling and segmenting sequential data and was first introduced by Lafferty et al. (2001). It 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.

The marginal probability specifies the model's confidence in each label of an input sequence, without the regard of the outcome other variables and can be used to measure a models' confidence in its predicted labelling (Sutton and McCallum, 2012). This can be computed through Constrained Forward-Backward, described in Culotta and McCallum (2004).

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

3.3. Unigram Tagger

The unigram tagger is used as a baseline and is a vocabulary based tagger. During training, the tagger adds every token to its vocabulary together with its label. The system predicts labels based on which label a token has been assigned to the most often. If the word has not been sen before it is tagged as not being personally identifiable information.

The concept is the same as NLTK's Unigram Tagger[5]. Although this is implemented in Scikit-learn.

The model is used as a baseline to indicate how a simple method would work and to enable comparison to other data sets. The tagger is meant to somewhat indicate the difficulty of the task.

3.3.1. Feature Set

The CRF is based on experiments with feature sets described in (Berg and Dalianis, 2019) except for a few altered regular expressions and no lemma or part of speech information used.

Each token is, first of all, itself a feature. Further lexical features used are parts of words. It also uses orthographic information to identify capitalisation, punctuation, pure numbers or a mix of numbers and letters. It also uses regular expressions to identify dates and phone numbers as well as binary dictionary features of whether a word exists in lists for first names, last names, hospitals and locations. Lexical information is also available for neighbouring words.

3.4. Bootstrapping Cycle

3.4.1. Input

There are three data sets used in this study: The *labelled set L*, the *unlabelled set* U and the *labelled evaluation set* E.

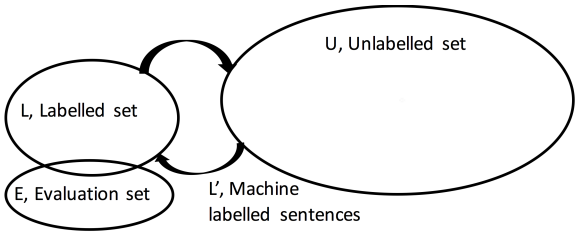

Figure 1: Figure over the three different data sets: The *labelled set L*, the *unlabelled set* U and the *labelled evaluation set* E.

The Stockholm EPR PHI Corpus is used for the labelled sets L and E, and a subset of the text in Health Bank is used for the unlabelled set U, see Figure 1. The extracted unlabelled dataset is 20 times larger than the manually labelled training data.

Initially, only L is included in the training set. The training set is used to train the CRF classifier. The trained classifier is used to obtain machine labelled sentences L' and the confidence of the label for each token in the sentence. If the label is above the set threshold for each token in a sentence, the token is added to L. The CRF is then retrained with the augmented set of both manually labelled and machine labelled sentences. This process of selection is repeated until all unlabelled sentences are processed, no more data meets the threshold or the F1 score has deteriorated for the development set for n times. n is set to 3 in this case.

3.4.2. Selection metric

The selection metric used to rank and select classified unlabelled instances for the next iteration is crucial to the performance of self-training (Wang et al., 2008).

The method considers the uncertainty of the estimated labels of the unlabelled data. The method assumes that the labels with high conditional scores are more likely to be true than lower scores. The added data is selected based on two properties: High marginal scores for the chosen label combined with low marginal scores for the other labels. The confidence threshold is decided based on the marginal scores for the development set for each iteration. Different marginal score thresholds are set for non-identifiable and identifiable tokens. Since tokens belonging to identifiable classes are a minority in the dataset. The inclusion of falsely labelled non identifying tokens are assumed to negatively affect recall, while the inclusion of tokens falsely labelled as identifying is assumed to primarily affect precision. The identifiable classes is a minority, and therefore likely to have lower confidence. As the focus is on finding identifying information, the selection threshold for these classes is therefore set separately from the threshold for the nonsensitive information.

3.4.3. Simple self-training method

The self-training method with selection, S, is also compared to a self-training method without selection, NS. All automatically labelled data is included.

[4]Linear-chain CRF, `https://sklearn-crfsuite.readthedocs.io/en/latest/`

[5]NLTK, Natural Language Toolkit, `https://www.nltk.org`

3.5. Evaluation

The models are evaluated as a named entity recognition task with micro-averaged recall, precision and F1 score. The recall is the ratio of correctly identified positive instances to all the instances in the actual class. Precision is the ratio of correctly identified positive instances to all the positive instances. F1 is the weighted average of recall and precision. Additionally, leakage scores are calculated. While recall correlates with exposure of identifiable information, not all recall errors cause this (Hirschman and Aberdeen, 2010). The leakage measurement is used to measure the number of identifiable terms not redacted in the de-identification process. The measurement for leakage is the ratio of identifiable information identified as non-identifiable to all the identifiable data. The leakage measurement used does not take into account the specificity and privacy risk of the leak. The leak of a patient's phone number has the same weight as a leak of "," in *"Neurology Unit, Huddinge"*.

4. Result

As a first step, the training data was iteratively added 10% at a time and evaluated on. As seen in Figure 2, an increase in the recall can be seen when adding more manually labelled training data, but the precision decreases slightly after the seventh iteration. No changes can be seen for *Age*, see Figure 3, after the second iteration where the first example of *Age* is seen. The age expressions in the training data only contain one example that does not follow the same format, and it seems that in an early stage learns that format. The model's ability to predict full dates does not either seem to considerably improve with time. Most dates follow the same pattern, which is distinctive from other number patterns. The ability to predict *Last Names* does also not change considerably for each iteration after a while. The biggest improvements can be seen for *Health Care Unit* and *Location*, where the model starts considerably lower than for the other classes. These are also the classes with the most variation in form and context. No organisations are found. This is likely because there is only one instance of an organisation in the training data.

Additionally, the Unigram Tagger is used as an additional baseline. As can be seen in Table 1, the precision is overall is 92.82%, while the recall is much lower at 44.55%. The precision of the Unigram Tagger is generally high. Since the tagger tags all previously unseen tokens as non-PHI, the lower recall is expected.

In the self-training experiment, the precision decreased, as can be seen in Table 3 compared to when training on only manually labelled data. The recall is increased slightly more than the precision is decreased, resulting in a higher F1 score. The biggest drops in precision can be seen for *Location* and *First Name*, with low increases in the recall in comparison. There is a slight drop in precision for *Health Care Unit* and *Phone Number* as well, but with increased recall. The recall increases for *Date Part*, *Last Name* as well, while it stays the same for *Full Date* and *Age*

Including automatically labelled data which meets the selection criteria performs significantly different than training on only manually labelled data according to McNemar's test (McNemar, 1947) with a Yates correction of 1.0. The

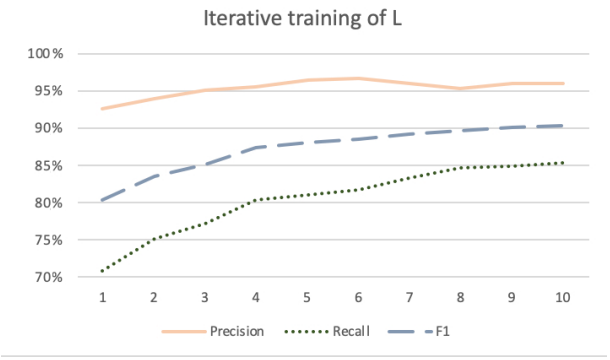

Figure 2: Micro-averaged results when iteratively adding 10% of L to training and then testing.

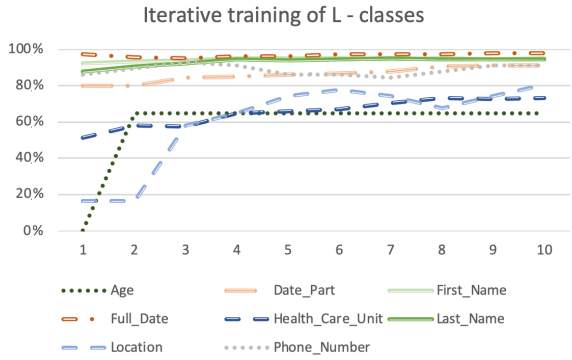

Figure 3: Recall for each class when iteratively adding 10% of L to training and then testing.

Yates correction is used since McNemar's test has an upwards bias, making results larger than they should be.

Comparing the self-trained model with a selection algorithm in Table 3 and a self-trained model with no selection algorithm presented in Table 4, the model with the selection criteria performs marginally better. The models perform similarly in regards to precision, while the model with the selection criteria gains higher recall. There is, however, no significant differences between using all the automatically labelled data, NS, or with a selection criterion.

As can be seen in Table 5, there is an increase of tokens labelled as identifiable between L and S. The biggest relative increase can be seen for the class *Location*, and the smallest if for *Phone Number*. The next smallest increase is for *Health Care Units*. Comparing the difference in labels added for S and NS, a conclusion can be drawn that these classes also are two classes linked to lower marginal probability scores.

5. Error analysis

While there is not a great difference in precision for most classes, there is a substantial decrease in precision for the identification of location in the self-training model. There is also a decrease for full dates.

While *Location* has low precision, no non-PHI is accidentally labelled as sensitive data. Despite the increase of examples labelled with Location, only one more *Location* is found. The issue is rather that health care units or part of

	Unigram Tagger	Baseline trained on labelled data L	Non-selection	Selection
Precision	92.82%	95.80%	94.14%	94.20%
Recall	44.55%	84.93%	88.16%	89.20%
F1	76.29	90.04	91.05	91.63
Leakage	33.90%	12.84%	9.40%	8.16%

Table 1: Comparison of precision, recall, F1 and leakage for all four models. The Unigram Tagger is described in subsection 3.3. Training Baseline is the model with only manually labelled data from L, with results further presented in Table 2. Selection is the semi-supervised model described in subsection 3.4, with results further presented in Table 3. Non-selection (NS) is the model where all automatically data is included and is further described in Table 4.

	Precision %	Recall %	F1
Age	100.00	64.52	78.43
Date Part	97.73	90.15	93.79
First Name	94.12	95.24	94.67
Full Date	95.83	97.87	96.84
Health Care Unit	95.01	72.88	82.49
Last Name	96.62	94.14	95.36
Location	78.12	80.65	79.37
Organisation	0.00	0.00	0.00
Phone Number	100.00	85.96	92.45
Overall	95.73	84.75	89.90

Table 2: Results from the token-based evaluation on the model based on the model with only manually labelled data, L.

	Precision %	Recall %	F1
Age	100.00	64.52	78.43
Date Part	97.85	94.93	96.36
First Name	89.94	95.83	92.80
Full Date	97.18	97.87	97.53
Health Care Unit	92.57	80.87	86.32
Last Name	97.39	95.60	96.49
Location	70.27	83.87	76.47
Organisation	0.00	0.00	0.00
Phone Number	92.86	91.23	92.04
Overall	94.20	89.20	91.63

Table 3: Results from the token-based evaluation on the model based on self-training with selection criteria, S.

	Precision %	Recall %	F1
Age	100.00	64.52	78.43
Date Part	96.54	91.64	94.03
First Name	93.02	95.24	94.12
Full Date	98.58	98.58	98.58
Health Care Unit	91.38	79.37	84.95
Last Name	97.05	96.34	96.69
Location	75.00	87.10	80.60
Organisation	0.00	0.00	0.00
Phone Number	94.55	91.23	92.86
Overall	94.14	88.16	91.05

Table 4: Results from the token-based evaluation on the model based on self-training without any selection criteria, NS.

Used sets	L	S	NS
Age	28	565	613
Date Part	328	5,984	6,725
First Name	589	7,779	10,416
Full Date	297	4,140	4,404
Health Care Unit	771	6,213	9,035
Last Name	627	7,995	11,184
Location	51	1,642	2,257
Phone Number	104	243	461
Overall	2,795	34,561	40,695

Table 5: The table presents the amount of annotated tokens in the original dataset (L), as well as the amount in the automatically labelled dataset with thresholds (S) and without thresholds.

health care units are mislabelled as locations. Since there are few *Location* annotations, small classification differences affects the precision and recall scores higher than for classes with more instances. The self-trained selection model classifies two more *Health Care Unit* tokens as *Location* and one more than the non-selection model. The misclassified health care units are referred to by their location, for example: "Schedule a return visit to Danderyd for discussion", where Danderyd refers to Danderyd's hospital in Danderyd. The non-selection model finds 55 more health care units than the baseline model, which may also affect this. In the automatically identified set, these types of references to Health Care Units are mislabelled as locations as well, which may cause further difficulty making a distinction between them for the model trained.

Out of the 138 identifiable tokens which are not found to be identifiable by the self-trained model with selection, 70% belong to the Health Care Unit class. This class is the most common class and the most variation with longer entities and short forms. It is also the class with the least clear boundaries. An example of this is for example "Urologen på St Göran", which could both be "The Urology department at St Göran" and "The urologist at St Göran". It is also somewhat unclear where the boundary for being identifying information should be set. In the automatically labelled for the selection model, unidentified Health Care Units are also included as non-PHI to a greater degree than for other classes.

The self-training model does not change the prediction ability. There is only one case in the original training data where an age n does not follow the format of "n-åring"

- so the model only learns to identify age by this.

6. Discussion

In de-identification research, the recall and precision are occasionally compared with each other and it is said that recall most often is more important than precision (Ferrández et al., 2012). Both these could, however, to some degree be altered, even if they are dependent on each other. This result does, however, to some degree indicate that it is possible to alter one without significantly harming the other. It is necessary to discuss how the goal of de-identification differs from the goal of named entity recognition in general. There is a need to further investigate how low-precision de-identification affects text mining research. Obeid et al. (2019) showed that de-identification with a system with demonstrated precision of 93% and a recall of 76% does not impact other text mining tasks. If the effect is low, it is an argument that techniques for increasing recall and sacrificing precision might be viable in situations where the desired recall is not fully met. Examples of this may not only be methods like the one proposed in this paper, but ensemble methods favouring recall over precision as well.

There have been previous discussions of how much scrubbing is enough, and where an accuracy of 95% is discussed (Stubbs et al., 2017). There is, however, still a need for some other type of risk assessment that takes into account the specificity and uniqueness of the information that is not found. All personally identifiable information is not equally identifying. A system which finds and masks 95% of the identifiable information but leaks the phone number of one patient's husband may breach patient privacy to a greater degree than a system that finds only finds 85% of the identifiable information, but where the information not found consists of general health care units, ages under 90 and date parts.

The use of self-training with unlabelled data does improve the system compared to using either a smaller amount of manually annotated medical data or a lot of out of domain data, as was carried out in Berg and Dalianis (2019). This may be due to the amount of identifiable information in the automatically labelled dataset combined with the domain similarities, or that the automatically labelled data complies with the guidelines in the Stockholm EPR PHI dataset to a greater degree than the general text source or smaller domain data.

7. Conclusion and Future Research

In this paper the research question addressed was whether it was possible to use semi-supervised learning to obtain more high-quality training data. The answer is: Yes, adding automatically labelled data based increases the model's ability to correctly identify identifiable information, without substantially affecting the precision. No difference could however be seen between including automatically labelled data in the training data and iteratively adding based on marginal probability scores. This paper also wants to highlight the need for measurements and techniques adapted for de-identification, beyond general techniques for named entity recognition.

We presented a method to machine annotate 40,000 tokens and preserve the high-quality of the produced data to be used as new training data. To manually annotate 40,000 tokens takes from 60 to 180 hours depending on if one has access to a pre-annotation method or not. It is a question of future research to investigate the quality of the machine annotated data and how it compares to human labelled data. The method adds a large amount of additional data, without considering which data is the most useful. In future work, investigating which data is necessary to include and not might prove important. Another possible direction would be to investigate other methods which do not require annotated data, including other semi-supervised techniques as co-training, or utilising transfer learning and word embeddings.

References

Berg, H. and Dalianis, H. (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*.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Boström, H. and Dalianis, H. (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.

Culotta, A. and McCallum, A. (2004). Confidence estimation for information extraction. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 109–112. Association for Computational Linguistics.

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

Dalianis, H., Henriksson, A., Kvist, M., Velupillai, S., and Weegar, R. (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.

Dasgupta, S., Littman, M. L., and McAllester, D. A. (2002). Pac generalization bounds for co-training. In *Advances in neural information processing systems*, pages 375–382.

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Ferrández, Ó., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H., and Meystre, S. M. (2012). Generalizability and comparison of automatic clinical text de-identification methods and resources. In *AMIA Annual Symposium Proceedings*, volume 2012, page 199. American Medical Informatics Association.

Haertel, R., Ringger, E., Seppi, K., Carroll, J., and McClanahan, P. (2008). Assessing the costs of sampling methods in active learning for annotation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 65–68. Association for Computational Linguistics.

Hanauer, D., Aberdeen, J., Bayer, S., Wellner, B., Clark, C., Zheng, K., and Hirschman, L. (2013). Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. *International Journal of Medical Informatics*, 82(9):821–831.

HIPAA. (2003). Health Insurance Portability and Accountability Act, U.S. Department of Health and Human Services. Accessed 2019-06-17.

Hirschman, L. and Aberdeen, J. (2010). Measuring risk and information preservation: toward new metrics for de-identification of clinical texts. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 72–75. Association for Computational Linguistics.

Kozareva, Z., Bonev, B., and Montoyo, A. (2005). Self-training and co-training applied to spanish named entity recognition. In *Mexican International conference on Artificial Intelligence*, pages 770–779. Springer.

Lafferty, J., McCallum, A., and Pereira, F. (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.

Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., and Solti, I. (2013). Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association*, 21(3):406–413.

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159. Association for Computational Linguistics.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Meystre, S. M. (2015). De-identification of Unstructured Clinical Data for Patient Privacy Protection. In *Medical Data Privacy Handbook*, pages 697–716. Springer.

Neamatullah, I., Douglass, M. M., Li-wei, H. L., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):1.

Nigam, K. and Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Cikm*, volume 5, page 3.

Obeid, J. S., Heider, P. M., Weeda, E. R., Matuskowitz, A. J., Carr, C. M., Gagnon, K., Crawford, T., and Meystre, S. M. (2019). Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in Health Technology and Informatics*, 264:283.

Rehbein, I., Ruppenhofer, J., and Sporleder, C. (2009). Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 19–26.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models. *WACV/MOTION*, 2.

Stubbs, A., Kotfila, C., and Uzuner, Ö. (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.

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

Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Velupillai, S., Dalianis, H., Hassel, M., and Nilsson, G. H. (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.

Wang, B., Spencer, B., Ling, C. X., and Zhang, H. (2008). Semi-supervised self-training for sentence subjectivity classification. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 344–355. Springer.

Wu, D., Lee, W. S., Ye, N., and Chieu, H. L. (2009). Domain adaptive bootstrapping for named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1523–1532. Association for Computational Linguistics.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Zhou, Y., Kantarcioglu, M., and Thuraisingham, B. (2012). Self-training with selection-by-rejection. In *2012 IEEE 12th international conference on data mining*, pages 795–803. IEEE.