Improving Named Entity Recognition and Classification in Class Imbalanced Swedish Electronic Patient Records through Resampling

Mila Grancharova
mila@dsv.su.se

Hanna Berg
hanna.berg@dsv.su.se

Hercules Dalianis
hercules@dsv.su.se

Department of Computer and Systems Sciences
Stockholm University
Kista, Sweden

Abstract

A key step in the de-identification of sensitive information in natural language text is the detection and identification of sensitive entities through Named Entity Recognition and Classification (NERC). Natural language data is often class imbalanced in two ways. First, with respect to the majority negative class consisting of all tokens other than named entities, and second, between the different classes of named entities. NERC of class imbalanced data often suffers in recall. This is an issue in de-identification systems where recall is essential in ensuring protection of sensitive information.

This study attempts to improve NERC, with focus on improving recall, in Swedish electronic patient records through resampling strategies involving negative class undersampling and minority class oversampling. The methods are evaluated in two NERC models based on machine learning methods. In both models, an increase in recall is achieved through undersampling, oversampling and combinations thereof. Undersampling, however, has negative effects on precision.

1 Introduction

Electronic patient records (EPR), also called clinical text, contain valuable information about patients' symptoms, diagnoses, treatments and treatment outcomes. Advancements in natural language processing and machine learning have made it possible to use large amounts of clinical text to create tools that assist physicians and medical researchers in detecting early symptoms of disorders, predicting adverse effects of treatments, etc, see Chapter 10 in (Dalianis, 2018). However, clinical text contains information that can reveal the identity of patients and other mentioned individuals, so called Protected Health Information (PHI). Here, PHI refers only to the named entities which may reveal a person's identity, such as name, age and location. In this sense, detecting and identifying the PHI before obscuring them is a Named Entity Recognition and Classification (NERC) problem.

Imbalance in the number of instances from different classes can have negative effects on classification through machine learning methods. In NERC, there are two types of class imbalance problems. The first is that in most sentences, the majority of the tokens do not belong to a named entity. In the commonly used data set Conll 2003, (Tjong Kim Sang and De Meulder, 2003), 16% of all tokens are part of a named entity. In clinical text, only an estimated 1-3% of all tokens are part of a PHI entity (Japkowicz, 2000; Velupillai et al., 2009; Buda et al., 2018). This means that there is a large majority class of negative samples, often annotated 'Other' or 'O'. Secondly, imbalance can exist between the various named entity classes.

Resampling techniques that may be used for balancing data include various forms of majority class undersampling and minority class oversampling. In the case of NERC, undersampling can be used to reduce the negative class. This task is non-trivial due to the syntactic relation between words in text, meaning that removing negative samples at random may eliminate useful contextual information.

Oversampling can be used to improve the balance between the positive classes. A common oversampling technique is random oversampling, where samples from one or more minority classes are duplicated at random. Research suggests that human learners use patterns of shared lexical contexts to form word categories when learning linguistic form-classes (Reeder et al., 2013; Clair et al., 2010). Since NERC is essentially the task of learning to separate such categories, this study explores weather it can be beneficial to combine random oversampling with surrogate generation to create

samples with shared lexical contexts. This is done by replacing the PHI entities in the duplicated sentences with surrogates after oversampling the minority classes.

In this study, two machine learning models for NERC are trained on data sets resampled with the above mentioned techniques, and combinations thereof, and tested on the original class-imbalanced data. A comparison is also made to models trained on the original data. The aim is to achieve high recall for better de-identification.

2 Related Research

On average, de-identification systems have better precision than recall (Stubbs et al., 2015, 2017; Marimon et al., 2019). When high recall is preferred over high precision to ensure the privacy of data subjects, this may be an issue. Gardner et al. (2010) suggest a window undersampling technique to compensate this, where only negative samples within a specified distance to positive samples are use to train the NERC system. Another undersampling method for NERC is the balanced undersampling method proposed by Akkasi et al. (2018). The method consist of attaining a pre-defined positive to negative sample ratio on the sentence level.

Berg and Dalianis (2020) attempt to improve NERC performance of Swedish clinical text by extending an annotated data set with data generated using a semi-supervised learning method. The aim is to increase recall without sacrificing precision. The authors report significant increase in recall and some decrease in precision. In a recent study on English clinical text de-identification, Yue and Zhou (2020) propose a method for PHI augmentation and context augmentation with the purpose of improving generalisation in NERC models.

3 Data and Methods

This chapter describes the data set, see Section 3.1, and methods, see Section 3.2, used to generate the results presented in this paper.

3.1 Data

The original data[1] used in this study is Stockholm EPR PHI Corpus, consisting of 200,000 tokens. Its annotation is described in Velupillai et al. (2009). The data was refined and used in the first de-identification experiment described in Dalianis and Velupillai (2010). Table 1 presents the distribution of the annotated PHI classes. The corpus is imbalanced in two ways. First, barely 3% of the tokens belong to PHI classes. Secondly, there is imbalance between the nine PHI classes.

PHI Class	Instances
First Name	923
Last Name	931
Phone Number	137
Age	55
Full Date	457
Date Part	709
Health Care Unit	1,414
Location	95
Organisation	43
Total	4,764

Table 1: Stockholm EPR PHI Corpus.

3.2 Methods

This study explores several resampling methods with the aim of reducing the negative class, thus the samples not belonging to a PHI class, and improving the balance between the positive classes. The resampling methods and resulting data sets are described in Section 3.2.1.

The resampling methods are evaluated through NERC on previously unseen data with models trained on the different data sets. The NERC models used are described in Section 3.2.2.

3.2.1 Resampling Methods

First, 20% of the original data, selected at random, is held out for testing. All models are evaluated on this test set. The remaining 80% are used to create several resampled training sets.

In an effort to reduce the number of negative samples, the balanced undersampling method proposed by Akkasi et al. (2018) is used. The idea behind the method is to select which negative samples to remove by considering their position with respect to positive samples. The method works by defining an undersampling parameter R, which should be attained in each individual sentence. For each sentence, if the positive to negative sample ratio, see Equation 1, is R or greater, no tokens are

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

removed. If not, the negative tokens furthest from a positive sample are removed iteratively until a ratio of R is attained or surpassed.

$$\text{ratio} = \frac{\sum \text{PHI tokens}}{\sum \text{other tokens}} \quad (1)$$

The value of R is found through experiments where the mean NERC performance is found on the training set undersampled with different values of R. Figure 1 presents the results of these experiments as a mean of all PHI classes. We see that the precision increases for a decreasing value of R until $R = 0.05$ and then begins to decrease. Thus, R is set to 0.05. Note that undersampling does not change the distribution of PHI classes since no positive samples are removed.

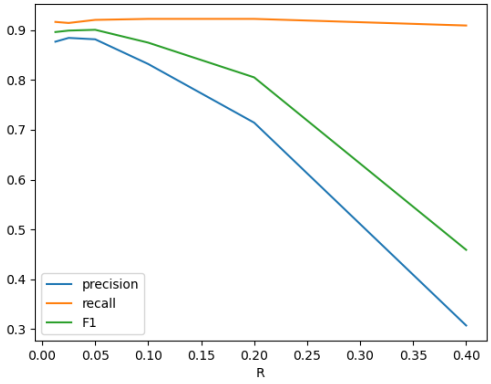

Figure 1: Mean NERC performance for different values of the undersampling parameter R.

When it comes to balancing the PHI classes, two approaches are examined. The first is classical random oversampling of four minority classes, namely *Age, Phone Number, Location* and *Organisation*. Sentences containing samples of these classes are duplicated at random until each class contains at least 370 samples. Sentences containing samples of the majority classes are excluded from duplication. The class distribution of the training set before and after random oversampling is shown in Table 2.

The second oversampling approach involves replacing PHI samples in the duplicated sentences with surrogates. The surrogate generation is lexical, based on the collection of Swedish named entity lists used in Dalianis (2019). We call this method pseudonymised random oversampling. The class distribution in this data set is the same as in the randomly oversampled data set, see Table 2.

Finally, the undersampling and oversampling methods are combined. Note that there are two combined training sets, one with random oversampling and one with pseudonymised random oversampling. Again, the PHI class distribution of these data sets is shown in the last column of Table 2.

PHI Class	Original	Over-sampled
First Name	728	728
Last Name	718	718
Phone Number	116	427
Age	46	466
Full Date	374	374
Date Part	578	578
Health Care Unit	1,121	1,121
Location	76	545
Organisation	33	405
Total	3,790	5,362

Table 2: PHI class distribution of the training set before and after oversampling.

Table 3 holds the ratio of positive tokens for each of the resampled training sets, as well as the the original training set. Note that oversampling does not change the ratio significantly since both positive and negative tokens are duplicated at random.

Training data set	%
Original Stockholm EPR PHI	3
Undersampled	13
Oversampled	3
Oversampled Pseudo	3
Undersampled & Oversampled	12
Undersampled & Oversampled Pseudo	12

Table 3: Positive token ratio of the training sets.

3.2.2 NERC Models

Two NERC models based on machine learning methods are trained on each of the data sets described in Section 3.2.1. The first is based on the linear-chain Conditional Random Fields (CRF) model implemented in CRFSuite (Okazaki, 2007). The model uses lexical, orthographic, syntactic and dictionary features, selected through experiments described in Berg and Dalianis (2019).

The second model is based on a bi-directional Long Short-Term Memory (BiLSTM) neural network with a CRF layer, based on a customised Tensorflow implementation[2]. Similar architectures

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

have shown success in NERC tasks in recent studies (Le et al., 2020; Kim et al., 2020). In essence, the BiLSTM extracts context-based representations of each token and these are then decoded in a linear-chain CRF layer. The model is fitted with Word2vec word embeddings of a Swedish clinical corpus containing 200 million tokens, in addition to the training set described in section 3.2.1.

4 Results

NERC results of the models trained on the training sets described above are presented in Table 4, holding the CRF results, and Table 5, holding the BiLSTM-CRF results. The results are all based on the original test set and are presented as a mean of the precision, P, recall, R and F_1-score across the nine PHI classes.

Training data set	P	R	$\mathbf{F}_1$
Original	0.9516	0.8890	0.9192
Undersampled	0.8819	**0.9209**	0.9010
Oversampled	0.9611	0.9147	**0.9373**
Over. Pseudo	**0.9661**	0.9085	0.9364
Und.&Over.	0.8843	0.9188	0.9012
Und.&Over. Pseu.	0.8728	0.9168	0.8942

Table 4: Mean CRF results across all PHI classes.

Training data set	P	R	$\mathbf{F}_1$
Original	0.8946	0.8650	0.8795
Undersampled	0.8060	0.8713	0.8374
Oversampled	**0.9067**	0.8790	**0.8926**
Over. Pseudo	0.8901	**0.8828**	0.8864
Und. & Over.	0.7193	0.8650	0.7855
Und.& Over. Pseu.	0.7300	0.8713	0.7944

Table 5: Mean BiLSTM-CRF results across all PHI classes.

5 Discussion and Conclusion

The results show that both undersampling the negative class and oversampling the minority positive classes can improve recall in NERC for class imbalanced data, see Table 4 and Table 5. In both NERC models, all tested resampling methods either improve recall on real data or leave it unchanged. The greatest improvement in recall comes from oversampling minority classes, with only slight difference between random oversampling and pseudonymised random oversampling, see Table 4 and Table 5. The results presented here surpass the recall of 89.20% reported by Berg and Dalianis (2020) on the same corpus. Since high recall is prioritised in a PHI de-identification system, these results show promise and suggest the methods should be explored and improved further.

Precision, on the other hand, does not seem to benefit from the resampling methods tested. In both NERC models, precision suffers significantly from undersampling, sometimes lowering the score by more than 10 percentage points. A reason for the decrease might be that the models have been trained on higher ratios of positive samples than are present in the test data, see Table 3, causing them to tend towards labelling more samples as positive, including negative samples. While sub-optimal, low precision is not a problem from a pure de-identification perspective. Still, for a useful system, high precision may still remain important. Keeping precision high while improving recall would be of interest and should be studied further.

An alternative to resampling is to use class weights in training. In other words, modifying the loss function as to penalise incorrect classification of samples from minority classes more than that of samples from majority classes. Using class weights to handle the positive to negative token ratio could be beneficial since, unlike in undersampling, the structure of the text would be preserved. It is possible that the negative effects on precision caused by oversampling in both models, and in in particular in the BiLSTM, are due to the discrepancies between the structure of the training data and the structure of the test data. This suggests an attempt at improvement by using class weights instead of undersampling.

Future work also includes using this method to balance data sets of completely pseudonymised training data from original non-deidentified data. Such data can be used for training privacy preserved machine learning models.

Acknowledgments

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

References

Abbas Akkasi, Ekrem Varoğlu, and Nazife Dimililer. 2018. Balanced undersampling: a novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text. *Applied Intelligence*, 48(8):1965–1978.

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*, pages 8–15, Turku, Finland. Linköping Electronic Press.

Hanna Berg and Hercules Dalianis. 2020. A semi-supervised approach for de-identification of swedish clinical text. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille*, pages 4444–4450.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Michelle C St Clair, Padraic Monaghan, and Morten H Christiansen. 2010. Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, 116(3):341–360.

Hercules Dalianis. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer.

Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

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.

James Gardner, Li Xiong, Fusheng Wang, Andrew Post, Joel Saltz, and Tyrone Grandison. 2010. An evaluation of feature sets and sampling techniques for de-identification of medical records. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 183–190.

Nathalie Japkowicz. 2000. The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence*, volume 56. Citeseer.

Juae Kim, Youngjoong Ko, and Jungyun Se. 2020. Construction of Machine-Labeled Data for Improving Named Entity Recognition by Transfer Learning. In *IEEE Access*.

Ngoc C. Le, Ngoc-Yen Nguyen, and Anh-Duong Trinh. 2020. On the Vietnamese Name Entity Recognition: A Deep Learning Method Approach. In *IEEE Access*.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *IberLEF@ SEPLN, Sociedad Española para el Procesamiento del Lenguaje Natural*, pages 618–638.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Patricia A Reeder, Elissa L Newport, and Richard N Aslin. 2013. From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, 66(1):30–54.

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.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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.

Xiang Yue and Shuang Zhou. 2020. PHICON: Improving Generalization of Clinical Text De-identification Models via Data Augmentation. In *The 3rd Clinical Natural Language Processing Workshop At EMNLP 2020. November 19, 2020*.