

PHICON: Improving Generalization of Clinical Text De-identification Models via Data Augmentation

Xiang Yue

The Ohio State University
yue.149@osu.edu

Shuang Zhou

The Hong Kong Polytechnic University
shuang.zhou@connect.polyu.hk

Abstract

De-identification is the task of identifying protected health information (PHI) in the clinical text. Existing neural de-identification models often fail to generalize to a new dataset. We propose a simple yet effective data augmentation method PHICON to alleviate the generalization issue. PHICON consists of **PHI** augmentation and **Context** augmentation, which creates augmented training corpora by replacing PHI entities with named-entities sampled from external sources, and by changing background context with synonym replacement or random word insertion, respectively. Experimental results on the i2b2 2006 and 2014 de-identification challenge datasets show that PHICON can help three selected de-identification models boost F1-score (by at most 8.6%) on cross-dataset test. We also discuss how much augmentation to use and how each augmentation method influences the performance.¹

1 Introduction

Clinical text in electronic health records (EHRs) often contain sensitive information. In the United States, Health Insurance Portability and Accountability Act (HIPAA)² requires that protected health information (PHI) (e.g., name, street address, phone number) must be removed before EHRs are shared for secondary uses such as clinical research (Meystre et al., 2014).

The task of identifying and removing PHI from clinical texts is referred as de-identification. Although many neural de-identification models such as LSTM-based (Dernoncourt et al., 2017; Liu et al., 2017; Jiang et al., 2017; Khin et al., 2018) and BERT-based (Alsentzer et al., 2019; Tang et al., 2019) have achieved very promising performance, identifying PHI still remains challenging in the

real-world scenario: even well-trained models often *fail to generalize to a new dataset*. For example, we conduct cross-dataset test on i2b2 2006 and i2b2 2014 de-identification challenge datasets³ (i.e., train a widely-used de-identification model NeuroNER (Dernoncourt et al., 2017) on one dataset and test it on the other one). The result in Figure 1 shows that model’s performance (F1-score) on the new dataset decreases up to 33% compared to the original test set. The poor generalization issue on de-identification is also reported in previous studies (Stubbs et al., 2017; Yang et al., 2019; Johnson et al., 2020; Hartman et al., 2020).

To explore what factors lead to poor generalization, we sample some error examples and find that the model might focus too much on specific entities and does not really learn language patterns well. For example, in Figure 2, given a sentence “*She met Washington in the Ohio Hospital*”, the model tends to recognize the entity “*Washington*” as the “*Location*” instead of the “*Name*” if “*Washington*” appears as “*Location*” in the training many times. Such cases appear more frequently in a new testing set, thus leading to poor generalization.

To prevent the model overfitting on specific cases and encourage it to learn general language patterns, one possible way is to enlarge training data (Yang et al., 2019). However, clinical texts are usually difficult to obtain, not to mention the requirement of tremendous expert effort for annotations (Yue et al., 2020). To solve this, we introduce our data augmentation method PHICON, which consists of **PHI** augmentation and **Context** augmentation. Specifically, PHI augmentation replaces the original PHI entity in the training set with a same type named-entity sampled from external sources (such as Wikipedia). For example, in Figure 2, “*Ohio Hospital*” is replaced by an randomly-sampled “*Hospital*” entity “*Alaska Health Center*”. In terms of context aug-

¹Our code is available at: <https://github.com/betterzhou/PHICON>

²<http://www.hhs.gov/hipaa>

³<https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/>

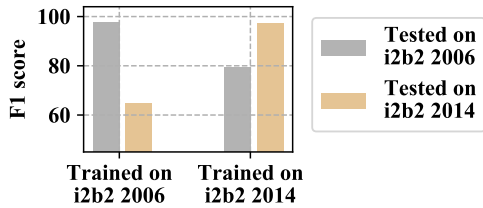


Figure 1: The result of cross-dataset test based on a base model (Dernoncourt et al., 2017). Performance on the new dataset drops up to 33% compared to the original test set, showing the model suffers from generalizability issue.

mentation, we randomly replace or insert some non-stop words (e.g., verb, adverb) in sentences to create new sentences as an example shown in Figure 2. The augmented data does not change the meaning of original sentences but increase the diversity of the data. It can better help the model to learn contextual patterns and prevent the model focusing on specific PHI entities. Data augmentation is widely used in many NLP tasks (Xie et al., 2017; Ratner et al., 2017; Kobayashi, 2018; Yu et al., 2018; Bodapati et al., 2019; Wei and Zou, 2019) to improve model’s robustness and generalizability. However, to the best of our knowledge, no work explores its potential in the clinical text de-identification task.

We test two LSTM-based models: NeuroNER (Dernoncourt et al., 2017), DeepAffix (Yadav et al., 2018) and one BERT-based (Devlin et al., 2019) model: ClinicalBERT (Alsentzer et al., 2019) with our PHICON. Cross-dataset evaluations on i2b2 2006 dataset and i2b2 2014 dataset show that PHICON can boost the models’ generalization performance up to 8.6% in terms of F1-score. We also discuss how much augmentation we need and conduct the ablation study to explore the effect of PHI augmentation and context augmentation. To summarize, our PHICON is simple yet effective and can be used together with any existing machine learning-based de-identification systems to improve their generalizability on new datasets.

2 PHICON

To understand what factors lead to the poor generalization, we check some error examples and find that most of the PHI entities in these error examples do not appear in training set or appear as a different PHI type (e.g., Washington [Name v.s. Location]). We argue that neural models might focus on too

Original Sentence		
She met Washington [NAME] in the Ohio Hospital [HOSPITAL]		
PHI Aug		
She met William [NAME] in the Alaska Health Center [HOSPITAL]		
Context Aug	SR	She saw Washington [NAME] in the Ohio Hospital [HOSPITAL]
	RI	She happily met Washington [NAME] in the Ohio Hospital [HOSPITAL]

Figure 2: Toy examples of our PHICON data augmentation. SR: synonym replacement. RI: random insertion.

much on specific entities (e.g., recognizing “Washington” as “Location”) but fail to learn general language patterns (e.g., “met” is not usually followed by a “Location” entity but a “Name” entity instead). Consequently, such unseen or Out-Of-Vocabulary PHI entities might be hard to be identified correctly, thus leading to lower performance. To help models better identify these unseen PHI entities, we may encourage models to learn contextual patterns or linguistic characteristics and prevent models focusing too much on specific PHI tokens.

PHI Augmentation. To achieve this goal, we first introduce PHI augmentation: create more training corpora by replacing original PHI entities in the sentence with other named-entities of the same PHI type. For example, in Figure 2, “Washington” is replaced by a randomly-sampled Name entity “William” and “Ohio Hospital” is replaced by an randomly-sampled Hospital entity “Alaska Health Center”.

We construct 11 candidate lists for sampling different PHI types. The lists are either obtained by scraping the online web sources (e.g., Wikipedia Lists) or by randomly generating based on pre-defined regular expressions (the number and the source of each candidate list is shown in Table 1).

Context Augmentation. To further help models focus on contextual patterns and reduce overfitting, inspired by previous work (Wei and Zou, 2019), we leverage two text editing techniques: synonym replacement (SR) and random insertion (RI) to modify background context for data augmentation (examples are shown in Figure 2). Specifically, SR is implemented by finding four types of non-stopping words (adjectives, verbs, adverbs and nouns) in sentences, and then replacing them with synonyms from WordNet (Fellbaum and Miller, 1998). RI is implemented by inserting random adverbs in front of verbs and adjectives in sentences, as well as inserting random adjectives in front of nouns in sentences.

Scraped from the Web					
PHI Type	Number	Source			
Organization	1,300	https://en.wikipedia.org/wiki/Category:Lists_of_organizations			
Hospital	5,400	https://en.wikipedia.org/wiki/Lists_of_hospitals_in_the_United_States https://www.hospitalsafetygrade.org/all-hospitals			
Location	27,500	https://en.wikipedia.org/wiki/List_of_Main_Street_Programs_in_the_United_States https://en.wikipedia.org/wiki/List_of_United_States_cities_by_area https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population			
Patient	14,900	https://en.wikipedia.org/wiki/List_of_most_popular_given_names			
Doctor	18,000	https://en.wikipedia.org/wiki/List_of_most_common_surnames_in_North_America			
Randomly Generated by Python scripts based on Regular Expressions					
ID	20,000	Username	3,000	Zip	4,000
Date	32,900	Phone	21,000	Medical Record	4,900

Table 1: The named-entity lists used for PHI augmentation, which are scraped from the Web or randomly generated.

		i2b2 2006	i2b2 2014	# PHI of each type	i2b2 2006			i2b2 2014		
					Train	Dev	Test	Train	Dev	Test
#notes	Train	622	912	CONTACT	159	32	41	394	31	96
	Dev	90	132	DATE	4887	649	1562	9102	974	2268
	Test	177	260	ID	3399	527	883	1000	166	312
	Total	889	1304	LOCATION	1761	252	648	3161	433	919
#avg tokens / note		631.7	810.8	NAME	3163	452	1064	5156	745	1439
#avg PHI / note		21.9	20.1	Total	13369	1912	4198	18813	2349	5034

Table 2: Statistics of the i2b2 2006 and 2014 datasets.

For each sentence containing PHI entities in the corpus, we can apply both PHI augmentation and Context augmentation to obtain the augmented data D_{aug} . We can run α times (by setting different random seeds) to obtain different sizes of augmented data (e.g., $\alpha = 2$ means augmenting the original dataset twice). Though with the α increases, we can obtain larger augmented training corpora, it may also bring much noise. We recommend a small value for α (See more discussions in Section 4.2). Then we merge the D_{aug} with the original dataset D to form the final dataset D_{new} for training: $D_{new} = D \cup \alpha D_{aug}$.

In summary, PHICON can significantly increase the diversity of training data without involving more labeling efforts. The augmented data can increase data diversity and enrich contextual patterns, which could prevent the model focusing too much on specific PHI entities and encourage it to learn general language patterns.

3 Experimental Setup

3.1 Datasets

We adopt two widely-used de-identification datasets: i2b2 2006 dataset and i2b2 2014 dataset,

and split them into training, validation and testing set with proportion of 7:1:2, based on notes number. We remove low frequency (occur less than 20 times) PHI types from the datasets. To avoid PHI inconsistency between the two datasets, we map and merge some fine-grained level PHI types into a coarse-grained level type, and finally preserve five PHI categories: Name (Doctor, Patient, Username), Location (Hospital, Location, Zip, Organization), Date, ID (ID, Medical Record), Contact (Phone). The statistics of the datasets are shown in Table 2.

3.2 Setup

Base Models. We select two LSTM-based models: NeuroNER (Dernoncourt et al., 2017)⁴, DeepAffix (Yadav et al., 2018)⁵ and one BERT model: ClinicalBERT (Alsentzer et al., 2019)⁶. All hyperparameters are kept the same as the original papers. **Evaluation.** To evaluate models' generalizability, we use the cross-dataset test on the two i2b2 challenge datasets: (1) Train the model on i2b2 2006 training set, and test on the whole i2b2 2014 dataset (Train + Dev + Test) (abbreviated

⁴<https://github.com/Franck-Dernoncourt/NeuroNER>

⁵https://github.com/vikas95/Pref_Suff_Span_NN

⁶<https://github.com/EmilyAlsentzer/clinicalBERT>

as “2006→2014”) (2) Train the model on i2b2 2014 training set, and test on the whole i2b2 2006 dataset (Train + Dev + Test) (abbreviated as “2014→2006”). For all experiments, we average results from five runs. We follow [Dernoncourt et al. \(2017\)](#) and report the micro-F1 score on binary token level.

4 Results

4.1 Does PHICON improve generalization?

In our preliminary experiments, we find that poor generalization tends to be more severe when the training set size is small. Thus, we consider the following training set fractions (%): {20, 40, 60, 80, 100} and we set the augmentation factor $\alpha = 2$ considering both effectiveness and time-efficiency (See the influence of α in Section 4.2). Table 3 shows the overall results, and interesting findings include:

(1) PHICON improves the generalizability of each de-identification model under different training sizes consistently. The results are not surprising as both PHI augmentation and context augmentation increase linguistic richness and enable models to focus more on language patterns, so as to help to train more generalized models.

(2) In general, the performance boost is large when the training data size is relatively small. This is because PHICON plays larger role at the low-resource case as it can significantly increase data diversity, language patterns, and linguistic richness.

(3) The performance boost on the BERT-based model is less obvious than that on LSTM-based models. Since ClinicalBERT has already been pre-trained on large-scale corpus: MIMIC-III clinical notes ([Johnson et al., 2016](#)). It is reasonable that the augmented data does not lead to large boost on ClinicalBERT. But there is still significant boost when training data size is relatively small.

(4) The boost on the setting “2006→2014” is larger than that in the setting “2014→2006”. Because i2b2 2014 dataset has more data and more comprehensive PHI patterns than i2b2 2006 dataset. Data augmentation is usually more effective when the training set size is smaller ([Wei and Zou, 2019](#)).

Improvement for each PHI category. To further understand PHICON, we show the performance (“2014→2006”) of the base model NeuroNER and NeuroNER + PHICON on each category of PHI in Figure 3. Firstly, we can see that when the training data is relatively small (e.g., 20%), the

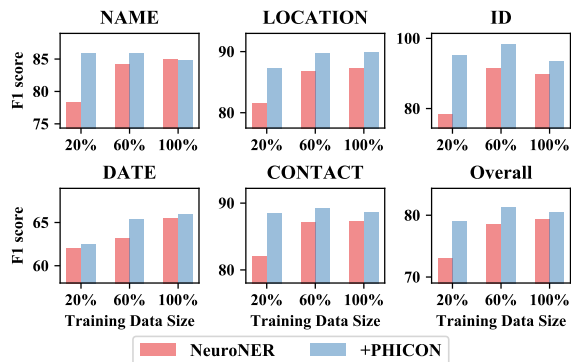


Figure 3: Performance of NeuroNER w/o and w/ PHICON on each PHI type (setting: 2014→2006)

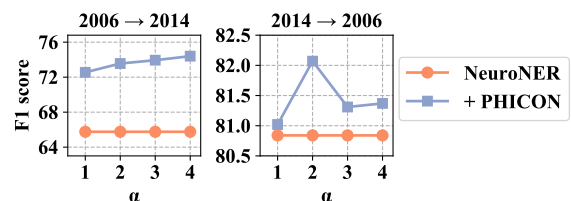


Figure 4: Data augmentation under different augmentation factors can boost model generalization. The left picture indicates that the model is trained on i2b2 2006 dataset and evaluated on i2b2 2014 validation set.

improvement on each PHI category is generally significant. With the training set size increases, the contribution of the augmented data becomes small. However, for the PHI categories that have less training data in the dataset (e.g., Location and ID; See Table 2), PHICON still contributes much improvement. Thus, we conclude that PHICON may be more helpful in the low-resource training data case.

4.2 How much augmentation?

In this section, we discuss the influence of the augmentation factor, α , on the cross-dataset test performance. In Figure 4, we report the performance on dev set based on the model NeuroNER for $\alpha = \{1, 2, 3, 4\}$. In the first setting (“2006→2014”), we can see the performance is steadily boosted with the increase of the factor α ; while in the second setting (“2014→2006”), the performance first goes up and then drops down. This difference might be caused by the data size of the two datasets (2014 dataset is larger). When the corpus is large, enlarging the augmentation factor might not lead to better performance, as the real data may have already covered very diverse language patterns. In addition, more augmented data might bring some

Trained on i2b2 2006, Tested on i2b2 2014

Model	Training Data Size				
	20%	40%	60%	80%	100%
NeuroNER (Dernoncourt et al., 2017)	0.5990	0.6021	0.6364	0.6436	0.6482
+ PHICON	0.6670	0.6979	0.7025	0.7063	0.7166
DeepAffix (Yadav et al., 2018)	0.5590	0.5875	0.6069	0.5976	0.6118
+ PHICON	0.6699	0.6543	0.6905	0.7170	0.6982
ClinicalBERT (Alsentzer et al., 2019)	0.7055	0.7149	0.7351	0.7454	0.7519
+ PHICON	0.7310	0.7412	0.7500	0.7586	0.7569

Trained on i2b2 2014, Tested on i2b2 2006

Model	Training Data Size				
	20%	40%	60%	80%	100%
NeuroNER (Dernoncourt et al., 2017)	0.7303	0.7513	0.7864	0.7891	0.7936
+ PHICON	0.7911	0.7944	0.8135	0.8175	0.8051
DeepAffix (Yadav et al., 2018)	0.6950	0.7467	0.7852	0.7774	0.7736
+ PHICON	0.7523	0.7706	0.7919	0.7827	0.8085
ClinicalBERT (Alsentzer et al., 2019)	0.8989	0.9043	0.9030	0.9069	0.9123
+ PHICON	0.9004	0.9076	0.9059	0.9078	0.9145

Table 3: Cross-dataset test performance (micro-F1 score on binary token level) on two experiment settings for models with and without PHICON on different training set sizes. All the numbers are the average from 5 runs.

Model	2006 → 2014	2014 → 2006
NeuroNER	0.648	0.794
+ PHI Aug	0.670	0.804
+ Context Aug	0.659	0.803
+ PHICON	0.717	0.805

Table 4: Ablation study on PHICON. PHI augmentation and context augmentation contribute to the overall generalization boost.

noise, which could decrease the performance. In terms of time efficiency, when α is increased by 1, the training time would roughly double if we set the same epoch number. So considering effectiveness, efficiency and data size, we recommend to set α a relative small value (e.g., 2) in the real application.

4.3 Ablation Study

In this section, we perform an ablation study on PHICON based on NeuroNER to explore the effect of each component: PHI augmentation and context augmentation. Table 4 shows that the two components of PHICON both contribute to boosting model generalization. Performance boost from PHI augmentation is obvious than context augmentation, i.e., PHI augmentation plays a major role. When combining both, PHICON results in larger boost than each of them.

5 Conclusion

In this paper, we explore the generalization issue on clinical text de-identification task. We propose a data augmentation method named PHICON that augments both PHI and context to boost model generalization. The augmented data can increase data diversity and enrich contextual patterns in training data, which may prevent the model overfitting on specific PHI entities and encourage it to focus more on language patterns. Experimental results demonstrate that our PHICON can help improve models’ generalizability, especially in the low-resource training case (i.e., the size of the original training set is small). We also discuss how much augmentation to use and how each augmentation method influences the performance. In the future research, we will explore more advanced data augmentation techniques for improving the de-identification models’ generalization performance.

Acknowledgments

We thank Prof. Kwong-Sak LEUNG and Sunny Lai in The Chinese University of Hong Kong as well as anonymous reviewers for their helpful comments.

References

- Emily Alsentzer, John Murphy, William Boag, Weihung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Sravan Babu Bodapati, Hyokun Yun, and Yaser Al-Onaizan. 2019. [Robustness to capitalization errors in named entity recognition](#). In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 237–242. Association for Computational Linguistics.
- Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *JAMIA*, 24:596–606.
- J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- C Fellbaum and G Miller. 1998. *WordNet : an electronic lexical database*. MIT Press.
- Tzvika Hartman, Michael D. Howell, Jeff Dean, Shlomo Hoory, and Yossi Matias. 2020. Customization scenarios for de-identification of clinical notes. *BMC Medical Informatics and Decision Making*, 20(1).
- Zhipeng Jiang, Chao Zhao, Bin He, Yi Guan, and Jingchi Jiang. 2017. De-identification of medical records using conditional random fields and long short-term memory networks. *JBI*, 75S:S43–S53.
- Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. [Deidentification of free-text medical records using pre-trained bidirectional transformers](#). In *ACM CHIL '20: ACM Conference on Health, Inference, and Learning, Toronto, Ontario, Canada, April 2-4, 2020 [delayed]*, pages 214–221. ACM.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad M. Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.
- Kaung Khin, Philipp Burckhardt, and Rema Padman. 2018. A deep learning architecture for de-identification of patient notes: Implementation and evaluation. *ArXiv*, abs/1810.01570.
- Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *NAACL'18*, pages 452–457.
- Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *JBI*, 75S:S34–S42.
- Stéphane M Meystre, Óscar Ferrández, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2014. Text de-identification for privacy protection: a study of its impact on clinical text information content. *JBI*, 50:142–150.
- Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. 2017. Learning to compose domain-specific transformations for data augmentation. In *NeurIPS*, pages 3236–3246.
- Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *JBI*, 75S:S4–S18.
- Buzhou Tang, Dehuan Jiang, Qingcai Chen, Xiaolong Wang, Jun Yan, and Ying Shen. 2019. De-identification of clinical text via bi-lstm-crf with neural language models. *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2019*:857–863.
- Jason W. Wei and Kai Zou. 2019. [EDA: easy data augmentation techniques for boosting performance on text classification tasks](#). In *EMNLP-IJCNLP'19*, pages 6381–6387. Association for Computational Linguistics.
- Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *ICLR'17*.
- Vikas Yadav, Rebecca Sharp, and Steven Bethard. 2018. [Deep affix features improve neural named entity recognizers](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 167–172, New Orleans, Louisiana. Association for Computational Linguistics.
- Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(Suppl 5):232.
- Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *ICLR'18*.
- Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. [Clinical reading comprehension: A thorough analysis of the emrQA dataset](#). In *ACL'20*, pages 4474–4486, Online. Association for Computational Linguistics.