Published in final edited form as:

Proc (IEEE Int Conf Healthc Inform). 2019 June ; 2019: . doi:10.1109/ICHI.2019.8904544.

A Study of Deep Learning Methods for De-identification of Clinical Notes at Cross Institute Settings

Xi Yang,
Tianchen Lyu,
Chih-Yin Lee,
Jiang Bian,
William R. Hogan,
Yonghui Wu
Health Outcomes and Biomedical Informatics College of Medicine, University of Florida Gainesville, USA

Abstract

In this study, we examined a deep learning method for de-identification of clinical notes at UF Health under a cross-institute setting. We developed deep learning models using 2014 i2b2/UTHealth corpus and evaluated the performance using clinical notes collected from UF Health. We compared four pre-trained word embeddings, including two embeddings from the general domain and two embeddings from the clinical domain. We also explored linguistic features (i.e., word shape and part-of-speech) to further improve the performance of de-identification. The experimental results show that the performance of deep learning models trained using i2b2/UTHealth corpus significantly dropped (strict and relax F1 scores dropped from 0.9547 and 0.9646 to 0.8360 and 0.8870) when applied to another corpus from a different institution (UF Health). Linguistic features, including word shapes and part-of-speech, could further improve the performance of de-identification in cross-institute settings (improved to 0.8527 and 0.9052).

Keywords

De-identification; Natural Language Processing; Deep Learning

I. Introduction

The unstructured clinical text has been increasingly used in clinical and translational research as it contains detailed patient information that not readily available in structured medical codes. De-identification [1] is a critical technology to facilitate the use of clinical narratives while protecting patient privacy and confidentiality. The Health Insurance Portability and Accountability Act (HIPAA) “Safe Harbor” rules identified 18 Protected Health Information (PHI) to be removed to generate a de-identified copy of clinical data. As manually de-identification is often time-consuming and not applicable to large volumes

alexgre@ufl.edu .

of clinical text, researchers have developed natural language processing (NLP) methods to identify and remove PHIs from clinical notes automatically. Most existing studies approach the de-identification as a clinical named entity recognition (NER) [2] task, which is a standard clinical NLP task to identify medical concepts and determine their semantic categories. The clinical NLP community has organized several shared tasks [3]–[5] to assess the current clinical NLP systems on de-identification of clinical text. In this study, we examined a deep learning method, LSTM-CRFs, for de-identification of clinical notes at UF Health under a cross-institute setting. We developed deep learning-based de-identification models using the 2014 i2b2/UTHealth corpus [3] and evaluated the performance using clinical notes collected from UF Health.

II. material and methods

A. Data sets

In this study, we used clinical notes from the 2014 i2b2/UTHealth challenge and UF Health Integrated Data Repository (IDR). For cross-institute evaluation, we collected a total number of 4,996 clinical notes from the UF Health IDR. These clinical notes were from 97 patients and distributed in 39 different note types. We randomly selected 218 notes from the UF Health dataset using stratified sampling based on the note types. Two annotators (TL and CL) manually annotated the PHIs in the selected notes. To facilitate cross-institute analysis, we merged several rare PHIs for the annotation of UF Health corpus: (1) excluded the *days of week, seasons* and *holidays*, *state* and *country* as they are not required by HIPAA; (2) merged the *phone* and *fax* as PHONE; (3) combined *email*, *URL* and *IP Address* as WEB; (4) merged organization and hospital as INSTITUTE. We adjusted the PHI annotations in the 2014 i2b2/UTHealth corpus to make the annotations consistent. Table I shows the distribution of PHIs in i2b2/UTHealth corpus and UF Health corpus.

B. Word embeddings

In this study, we examined four different word embeddings trained with different algorithms and corpora. The two general domain-based embeddings are released by Google and Facebook. The GoogleNews-word2vec embeddings were developed by Google trained using the word2vec on the part of the Google news dataset [6] and the CommonCrawl-fastText embeddings were released by Facebook trained using the fastText [7] algorithm and the Common Crawl dataset [8]. The two clinical domain-based embeddings are created by our group based on the corpus consisted of all notes from the Medical Information Mart for Intensive Care III (MIMIC-III) database [9]. We generated the MIMIC-III-word2vec using the word2vec while developed the MIMIC-III-fastText embeddings using the fastText.

C. Experiments and Evaluation

We used an LSTM-CRFs model developed in our previous work [10] using Tensorflow [11]. We compared the performance of LSTM-CRFs models with or without linguistic features (i.e., word shapes and part-of-speech). For evaluation, we reported the micro-averaged strict and relax precision, recall, and F1-score.

III. Results

Two annotators annotated 3,216 PHIs from 218 UF Health notes with an inter-annotator agreement of 0.889 Cohen’s kappa. We fixed the discrepancies of annotations through group discussions. Table I shows the detailed number of PHIs for each category and compared with the i2b2/UTHealth corpus. The model trained with the CommonCrawl-fastText achieved the best strict and relax F1 scores of 0.9547 and 0.9646 respectively on the I2B2 corpus. When applying this model to UF Health data, it achieved strict and relax F1 scores of 0.8360 and 0.8870, respectively. After adding linguistic features, both strict and relax F1 scores were improved to 0.8527 and 0.9052. Table II compares the performances (micro-averaged strict and relax F1-scores) of LSTM-CRFs model on the I2B2 dataset and the UF Health dataset.

IV. Discussion and Conclusion

This study shows that it is necessary to customize the deep learning-based de-identification systems for cross-institute settings.

Acknowledgment

The research reported in this publication was supported by the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences under award number UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We would like to thank the i2b2 organizers for providing the annotated corpus and guidelines for de-identification. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

References

[1]. Meystre SM, Friedlin FJ, South BR, Shen S, and Samore MH, “Automatic de-identification of textual documents in the electronic health record: a review of recent research,” BMC Med. Res. Methodol, vol. 10, p. 70, Aug. 2010. [PubMed: 20678228]

[2]. Nadkarni PM, Ohno-Machado L, and Chapman WW, “Natural language processing: an introduction,” J. Am. Med. Inform. Assoc. JAMIA, vol. 18, no. 5, pp. 544–551, Oct. 2011. [PubMed: 21846786]

[3]. Stubbs A, Kotfila C, and Uzuner Ö, “Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1,” Proc. 2014 I2b2UTHealth Shar.-Tasks Workshop Chall. Nat. Lang. Process. Clin. Data, vol. 58, pp. S11–S19, Dec. 2015.

[4]. Stubbs A, Filannino M, and Uzuner Ö, “De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1,” J. Biomed. Inform, vol. 75S, pp. S4–S18, Nov. 2017. [PubMed: 28614702]

[5]. Uzuner O, Luo Y, and Szolovits P, “Evaluating the state-of-the-art in automatic de-identification,” J. Am. Med. Inform. Assoc. JAMIA, vol. 14, no. 5, pp. 550–563, Oct. 2007. [PubMed: 17600094]

[6]. Mikolov T, Sutskever I, Chen K, Corrado G, and Dean J, “Distributed Representations of Words and Phrases and Their Compositionality,” in Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, USA, 2013, pp. 3111–3119.

[7]. Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, and Mikolov T, “FastText.zip: Compressing text classification models,” ArXiv Prepr. ArXiv161203651, 2016.

[8]. Mikolov T, Grave E, Bojanowski P, Puhrsch C, and Joulin A, “Advances in Pre-Training Distributed Word Representations,” CoRR, vol. abs/1712.09405, 2017.

[9]. Johnson AEW et al. , "MIMIC-III, a freely accessible critical care database," Sci. Data, vol. 3, p. 160035, May 2016. [PubMed: 27219127]

[10]. Wu Y, Yang X, Bian J, Guo Y, Xu H, and Hogan W, "CombineFactual Medical Knowledge and DistributedWord Representation to ImproveClinical Named Entity Recognition," in AMIA 2018 Annual Symposium, in press.

[11]. Abadi Martín et al., TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015.

TABLE I.

PHI distributions in the 2014 i2b2/UTHealth corpus and UF Health corpus.

PHI Category	**Number of Annotations**		
	2014 i2b2/UTHealth		***UF Heath***
	Training	***Validation***	***Evaluation***
DATE	9,067	3,104	1,866
NAME	5,472	1,868	782
AGE	1,507	490	166
ID	1,142	364	138
PHONE	406	128	46
WEB	6	1	4
INSTITUTE	1,926	592	129
STREET	280	72	21
CITY	502	152	44
ZIP	276	76	20
Total	20,584	6,847	3,216

TABLE II.

performance of LSTM-CRFs with knowledge features as embeddings

Data set	Model	Performance					
		Strict			***Relax***		
		Precision	***Recall***	***F1 score***	***Precision***	***Recall***	***F1 score***
I2B2	LSTM-CRFs	0.9697	0.9401	0.9547	0.9797	0.9498	0.9646
UF Health	LSTM-CRFs	0.8666	0.8075	0.8360	0.9195	0.8568	0.8870
	LSTM-CRFs + Lexical	0.8831	0.8242	0.8527	0.9376	0.8751	0.9052