MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231242

A De-Identification Model for Korean Clinical Notes: Using Deep Learning Models

Junhyuk CHANG[a] , Jimyung PARK[a], Chungsoo KIM[a] and Rae Woong PARK[b,1]
[a] *Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Korea*
[b] *Department of Biomedical Informatics, Ajou University School of Medicine, Korea*
ORCiD ID: Junhyuk Chang https://orcid.org/0000-0001-9622-0232, Jimyung Park https://orcid.org/0000-0002-6998-2546, Chungsoo Kim https://orcid.org/0000-0003-1802-1777, Rae Woong Park https://orcid.org/0000-0003-4989-3287

Abstract. To extract information from free-text in clinical records, due to the patient's protected health information (PHI) in the records, pre-processing of de-identification is required. Therefore we aimed to identify PHI list and fine-tune the deep learning BERT model for developing de-identification model. The result of fine-tuning the model is strict F1 score of 0.924. Due to the convinced score, the model can be used for the development of a de-identification model.

Keywords. Electronic health record, natural language processing

1. Introduction

Free text in electronic health records (EHR) contains essential clinical information such as the patient's medical profiles, symptoms. However, due to the patient's protected health information (PHI) in the reports [1], free text requires an additional de-identification process before usage. Although in 2020, Korea Personal Information Protection Commission issued a Guideline for Pseudonymization Published and the importance of PHI de-identification is remarkably increased for using medical big data, there have been few prior studies on de-identification in Korea, and insufficient development conducted. Therefore, we aimed to identify PHI list and fine-tune BERT model for developing PHI de-identification tool.

2. Methods

Fig 1. describes the framework and workflow of this study. In this study, we used the Ajou University School of Medicine database that is converted into the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) format. We identified entities related to patient privacy from 2,000 randomly sampled notes and annotated notes using schema which is constructed of essential PHI entities: name,

[1] Corresponding Author: Rae Woong Park; 507, Hongjae hall, 206, Worldcup–ro, Yeongtong–gu, Suwon–si, Gyeonggi–do, 16499; email: veritas@ajou.ac.kr; Telephone: 031-219-4457; Fax: +82–31–219–4472; ORCID: 0000-0003-4989-3287

residence, and contact. We used the deep learning BERT model, which is known to have high performance, to aim for constructing a PHI de-identification model and used a batch size of 4, training steps of 500, a learning rate of 1e-5, and a training epoch of 7 for fine-tuning the model. After adjusting the model parameter, we used precision, recall, and F1 score to evaluate fine-tuning performance.

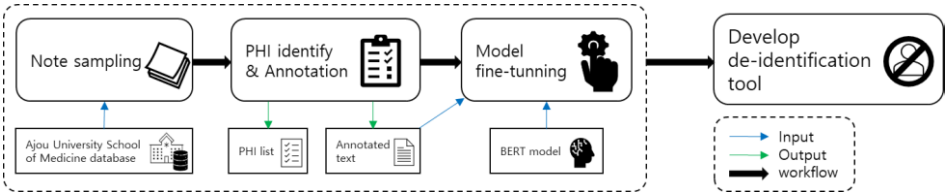

Figure 1. The workflow of a de-identification framework.

3. Results

A total of 4,511 entities are annotated from 2,000 notes and each number of entities are follows: name (n = 3,002), contact (n = 577), and residence (n = 932). Details of PHI entities are name of clinician and patient, contact number, residence of general pattern and birthplace.

Table 1 shows the evaluation results measured with the model and used categories. Fine-tuned BERT model had strict F1 score of 0.924 and relax F1 score of 0.948. C_number had the highest strict F1 score of 1.000 and R_general had the lowest strict F1 score of 0.739 among the categories. But the most essential privacy-related entities such as name of clinician and patient had high strict F1 scores of 0.948 and 0.981.

Table 1. Evaluation results of model fine-tuning.

Model or Category	Strict			Relax		
	Precision	Recall	F1 score	Precision	Recall	F1 score
BERT	**0.906**	**0.942**	**0.924**	**0.930**	**0.967**	**0.948**
N_clinician	0.930	0.967	0.948	0.945	0.972	0.958
N_patient	0.963	1.000	0.981	0.963	1.000	0.981
R_general	0.680	0.810	0.739	0.787	0.937	0.855
R_birthplace	0.857	0.750	0.800	1.000	0.875	0.933
*C_number	1.000	1.000	1.000	1.000	1.000	1.000

***C_: contact; N_: name of; R_: residence of**

4. Conclusions

In this study, we fine-tuned the model for PHI identification in Korean clinical reports. The fine-tuning score of the model came out convincing and considering the agglutinative language characteristics of the Korean language, it can be used for the development of a de-identification model.

References

[1] Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, Godtliebsen F. Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdisciplinary Reviews: Computational Statistics. 2021 Nov; (6):e1549, doi: doi.org/10.1002/wics.1549.