A de-identification model for clinical notes using deep learning: Application to Korean language

Junhyuk Chang Pharm.D.[1], Jimyung Park[2], Chungsoo Kim Pharm.D.[1], Rae Woong Park M.D., Ph.D[1, 3]

[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea
[2]Department of Biomedical Informatics, Columbia University, New York
[3]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea

Background

- The increasing use of electronic health records (EHR) has sparked interest in extracting information from free text within EHR
- Protected health information (PHI) in clinical reports requires de-identification before using the free text
- In South Korea, studies on PHI identification exist, but there is a lack of cross-institutional patient data applications
- Limited prior studies on PHI de-identification in medical big data have been conducted in Korea, resulting in inadequate development in this area
- Our objective was to identify the PHI list and fine-tune the BERT model for developing a PHI de-identification tool

Results

1. Annotated PHI entities

Name	Contact number	Residence
2,877	581	1,036

2. Conducted dictionaries

Name	Residence
3,112,431	19,523

3. Evaluation results of the fine-tuned model
- Calculate precision, recall, and relax F1 score
- Relax F1 score
  : I tag is considered the equivalent as B tag

Model and Category	Relax Precision	Recall	F1 score
BERT	**0.962**	**0.977**	**0.969**
Name	0.945	0.972	0.958
Residence	1.000	0.980	0.990
Contact	0.963	1.000	0.981
BERT +regex+dict	**0.961**	**0.985**	**0.973**
Name	0.946	0.984	0.965
Residence	0.981	1.000	0.990
Contact	1.000	0.980	0.990

Table 2. Sample of gold standard and predicted text

PHI entities	Gold standard	Predicted text
Name	consult O to O 김철수 B-NAME 교수님 O	consult O to O 김철수 B-NAME 교수님 O
Residence	Address O : O 경기 O 수원시 O 원천동 B-RESIDENCE 중부대로 I-RESIDENCE 123-45 I-RESIDENCE	Address O : O 경기 O 수원시 O 원천동 B-RESIDENCE 중부대로 B-RESIDENCE 123-45 B-RESIDENCE
Contact	전화번호 O 123-4567 B-CONTACT " O	전화번호 O 123-4567 B-CONTACT " O

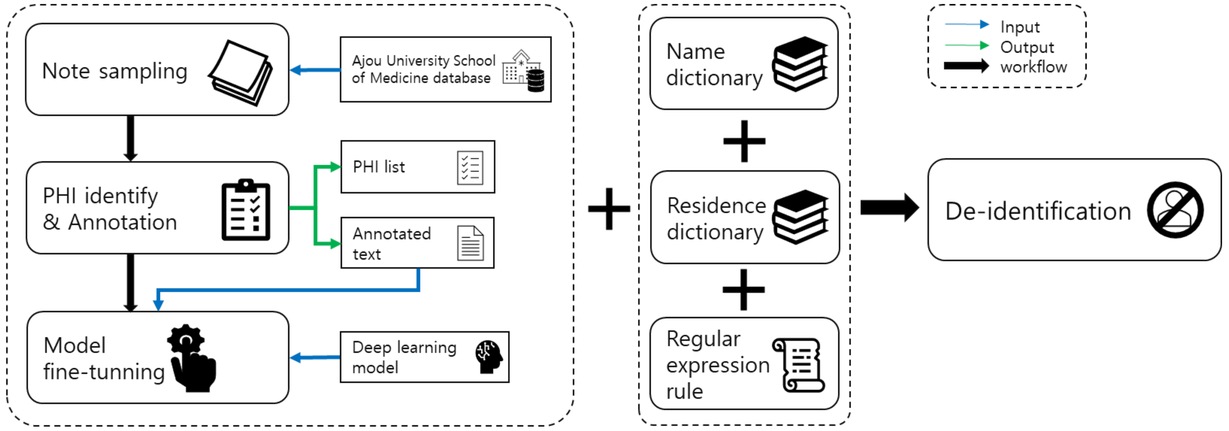
Conclusions

- In this study, we successfully fine-tuned a model for PHI identification in Korean clinical reports
- This model can serve as a foundation for developing a de-identification tool
- based on this model, in further study, the software application has to be developed and evaluated by external validation in various hospital database

Method

1. Data source
- AUSOM database
  (Ajou University School of Medicine)
- Contained ≥ 2.9 million patients in OMOP-CDM format
- Collected from Jan 1994 to Apr 2021

2. Framework and workflow



3. Annotated PHI
- Define PHI list and annotate
  ➢ Extracted 3,000 randomly sampled notes
  ➢ Identified PHI entities and annotated sampled notes using schema which is constructed of essential entities:
    ① Name
    ② Contact number
    ③ Residence

4. Conducted dictionary
  1) Name dictionary
: Contained about 3 million synthetic Korean names

  2) Residence dictionary
: Based on the law of the Three Revised Bills that Ease Regulations on the Use of Personal Information
: Contained administrative district residences that a sub-address of town/city and state

Acknowledgement

- This research was funded by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR16C0001), and a grant from the project for Infectious Disease Medical Safety, funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HG22C0024)