De-identification of Clinical Notes for Patients with Infectious Disease and Topic Modeling using Latent Dirichlet Allocation

Junhyuk Chang[1], Jimyung Park[1], Chungsoo Kim[1], Rae Woong Park[1, 2]
[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea;
[2]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea

Background

As the electronic health record (EHR) are widely used, secondary use of EHR data has been increased such as the development of predictive models and deploying syndromic surveillance system. Due to the coronavirus-19 global pandemic, infectious disease-related research and methodologies have been remarkably increased.

However, the essential clinical information such as the patient's medical profiles, disease symptoms, and treatment results are usually recorded in clinical reports with the form of free-text. Textual data can be extracted and processed through two distinct approaches: 1) manual chart review or 2) natural language processing (NLP). NLP is a computational analysis that can reduce the laborious burden of chart review, however, due to patients' protected health information (PHI) in the reports[1], clinical NLP requires an additional de-identification process before NLP processes.

Therefore, to identify the infectious disease-related features from the clinical reports, PHI should be identified and eliminated. In 1996, the US Department of Health and Human Services issued the Heath insurance portability and Accountability Act (HIPAA) Privacy Rules and defined 18 types of PHI and conducted research on the de-identification of PHI. In the Republic of Korea, although studies on the identification of PHI have been conducted, there are insufficient cases of application to patient data from other institutions.

In this study, we applied the NLP technique based on unsupervised learning at the word level to confirm the distribution of information on the clinical notes of infectious disease patients after de-identifying PHI through regular expression rules.

Methods

1. Data preparation

In this study, we used the Ajou University Medical Centre database that is converted into the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) format. The target population was defined as the patients who admitted to the hospital from January 2012 to December 2021 and diagnosed with any of infectious diseases within two days before and after the hospitalization. We used the Systematized Nomenclature of Medicine Clinical Terms code '40733004' and its sub-hierarchy codes for the infectious disease diagnosis. The admission notes of the target population were extracted and used in the study.

2. PHI identification and de-identification

We compared our corpus with the HIPAA PHI list to identify the potential PHI entity in Korean clinical reports. A thousand of admission notes were randomly selected from the corpus and manually compared with the HIPAA PHI list. The identified PHI entities were anonymized with the two approaches: 1) rule-based approach and 2) dictionary-based approach. Especially, since proper nouns (e.g., patient name, city, country, and hospital) should be considered precisely, we constructed dictionaries per each relevant PHI entity. To eliminate patients' names, we constructed a name dictionary containing 47,699 names made of combining ten first names of preference for each year from 1940 to 2019 and last names with the established name dictionary. We also extracted the names of country, hospital, city and state from public open data and constructed a separated dictionary to identify them. Regular expression rules are also constructed to identify other PHI patterns.

3. Feature identification using topic modeling

With the anonymized corpus, we used the latent Dirichlet allocation (LDA) model to identify the infectious disease related features. LDA is a notable unsupervised topic modeling method that can cluster the documents by semantic topics. Each topic (cluster) can demonstrate their belonging token; therefore, the users can infer the semantic meanings of the corpus. To decide the optimal number of topics, we used the perplexity score calculation algorithm[2] and semantic meanings of topics.

Results

1. Data preparation

A total of 44,415 patients were identified as the patients who were diagnosed with disorder due to infection within two days of admission, and their 61,379 admission notes were used in the study. We eliminated the numbers, the punctuations of admission notes for pre-processing for LDA application. Additionally, the authors decide that unit-related words (i.e., ul, tab, ml) and specialty-related words (i.e., gastrointestinal internal medicine, pulmonary internal medicine) were not meaningful, hence, added to the stopwords list.

2. PHI identification and de-identification

By comparing the HIPAA PHI with Korean clinical notes, we were able to identify overall 9 PHI entities that were recorded into 21 patterns (Table 1). The identified PHI entities were name, family relationship, contact, country name, state and city name, birthplace and residence, hospital room number, and profession. Of those, name, birthplace and residence, hospital room number related information were found the most various patterns. The names of patients, country, hospital, and region were anonymized with the in-house developed name dictionary. Meanwhile, the other patterns were anonymized with the defined regular expression rules.

3. Feature identification using topic modeling

According to the perplexity score algorithm, we found that number of 5 or 8 topics were optimal number of topics. However, the authors decided to review only number of topics with 5 or close number (i.e., 4 and 6) for the clear explanation of the semantic meanings. As a result, LDA with the four topics was identified and used in the study, which showed similar marginal topic distribution between each other topics (Figure 1).

Figure 2 shows the most frequently identified tokens per each topic. Circulatory system-related words (i.e., htn, caod, chest) were clustered in topic 1, and hepatobiliary system-related words (i.e., hcc, hepatitis, baraclude) were clustered in topic 2. Pediatric-related words (i.e., 환아로 [pediatric patient in Korean]), sepsis, uti, seizure, apn) were clustered in topic 3, and respiratory system-related words (i.e., pneumonia, cough, uri, dyspnea) were clustered in topic 4. The high-density words such as fever, pneumonia, cough, and uti suggest that infectious disease-related features can be extracted from clinical notes.

Conclusion

In this study, we defined PHI de-identification algorithm in Korean clinical reports and were able to apply it to the admission notes of patients diagnosed with infectious diseases. Furthermore, the infectious symptoms and signs in the corpus were identified using the LDA model. Further study using extracted symptoms of the infectious disease should be considered and the overall analytic pipeline using the NLP technique should be standardized for the distributed network research.

Acknowledgement

This work was supported by the Bio Industrial Strategic Technology Development Program (20003883, 20005021) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea), and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR16C0001). This research was also supported by a grant of 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: HG22C0024).

References

1. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records 2021;13:n/a.
2. Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In: Zaki MJ, Yu JX, Ravindran B, Pudi V, editors., Berlin, Heidelberg: Springer Berlin Heidelberg; 2010, p. 391–402.

Table 1. Twenty-one regular expression rules for de-identification of Korean clinical notes.

Patient of identifiers	Rules	Example
Name		
Patient and family	(1) Exact match using a list of names from name dictionary	James → ***

Relatives of patient	(2) Dr[.]\W{2,3}
	(3) R\d.\W{2,3}
	(4) pf[.]\W{2,3}	

Family relationship		
Relationship with patient	(5) 보호자[:punct:]\W{0,10}[:punct:]	Companion(father) → Companion(***)
	(6) Exact match using a list of family relationships from open data	daughter → ***
Contact		
Phone	(7) (전화번호|전화 *번 *호 *|번호)[0-9]{3}-[0-9]{3,4}-[0-9]{3,4}	123-4567-8901 → ***_****_****
	(8) (전화번호|전화 *번 *호 *|번호)[0-9]{3}-[0-9]{3,4}	123-4567 → ***_****_****
Location		
Country	(9) Exact match using a list of country names from open government database	Country name → ***
State and City	(10) Exact match using a list of state and city names from open database	State name → ***
Birthplace and residence	(11) (출생지\/거주지\:|출신지\s\/\s 거주지\s\:)\s{0,1}([가-힣]{2,5}|[가-힣]{2,5}\s[가-힣]{2,5})\s{0,1}\/{1}\s{0,1}[가-힣]{2,5}\s[가-힣]{2,5}\s	birthplace/residence : Seoul / Gyeonggi-do → birthplace/residence : ***
	(12) (출생지\/거주지\:|출신지\s\/\s 거주지\s\:)\s{0,1}([가-힣]{2,5}|[가-힣]{2,5}\s[가-힣]{2,5})\s{0,1}\/{1}\s{0,1}[가-힣]{2,5}	
	(13) [^\/\n]{0}거주지\s{0,1}\:\s([가-힣]{2,10}\,\s[가-힣]{2,10}|[가-힣]{2,10})	residence : Seoul → residence : ***
	(14) 거주지\s[가-힣]{2,10}	
Hospital	(15) Exact match using a list of hospital names from open government database and in-house database	John hospital → *** hospital
	(16) ([^-\s(]{0,}[^,타]병\s*원)	
	(17) [^\s]{0,}의료원	John medical center → ***medical center
	(18) [^\s]{0,}의료재단	John medical foundation → *** medical foundation
Hospital room number	(19) ([A-Z0-9]{1,3}W-[0-9]{1,2}-[0-9]{1,2}|[0-9A-Za-z]{1,2}[CU]{1,2}[0-9A-Za-z]-[0-9]{1,2}-[0-9]{1,2}|[ER]{1,2}-[0-9]{1,2}-)	ABCD-12-34 → ****-**-**
Family history		
Relationship with patient	(6) Exact match using a list of family relationships from open data	FHx Brother → FHx ***
Profession		
Profession	(20) 직업\:\s([가-힣]{2,10}\,\s[가-힣]{2,10}|[가-힣]{2,10})	Profession : engineer →profession : ***
	(21) 직업\s[가-힣]{2,10}	

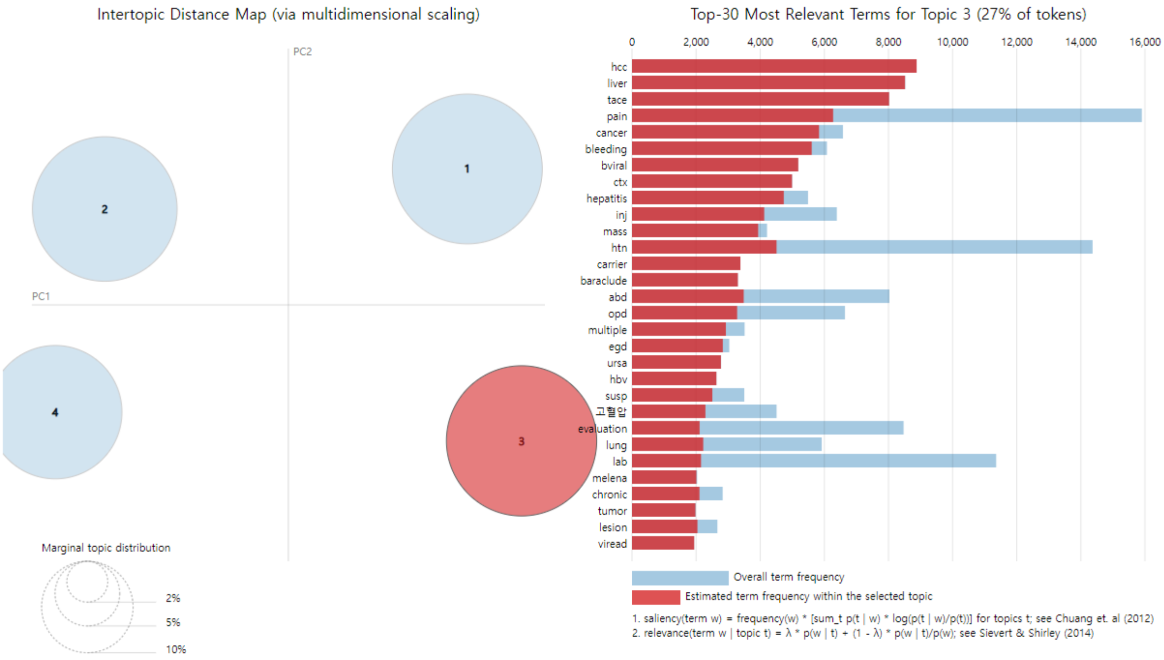

Figure 1. Topic distance map and terms for the topic 3.

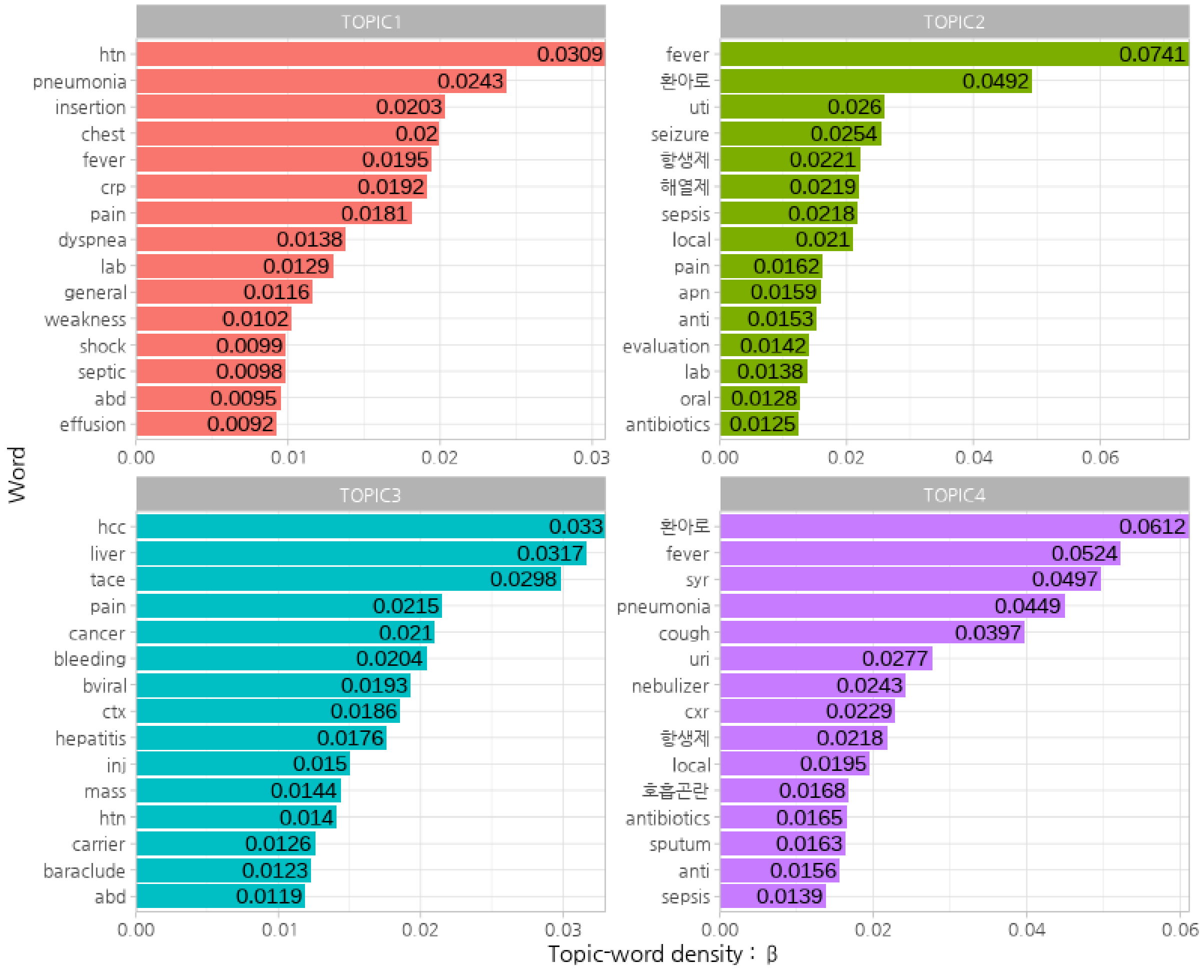

Figure 2. Word density plot for four topics. Each Korean words mean as follows: 환아로 (pediatric patient); 항생제 (antibiotics); 해열제 (antipyretics); 호흡곤란 (dyspnea).