Topic Modeling of Clinical Notes for Patients with Infectious Disease using Latent Dirichlet Allocation after Deidentification of Protected Health Information

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) is widely used, secondary use of EHR data has been increased. Due to coronavirus-19 global pandemic, researches on infectious disease using EHR has been remarkably increased.

However, essential clinical information such as patients' medical profiles, disease symptoms, and treatment results in clinical reports are usually recorded in the form of free-text, and these unstructured data are hard to use. These free-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 potentially protected health information (PHI) exists in the free-text[1], clinical NLP requires a prior de-identification process.

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 PHI identification studies have been conducted, there has been insufficient external validations.

In this study, we applied two approaches to de-identifying PHI and applied NLP techniques based on unsupervised learning at the word level to confirm the distribution of information on the clinical notes on infectious disease.

Methods

1. Data preparation

We used Ajou University Medical Centre database that is converted into the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) format. The target populations are defined with two inclusion criteria: 1) The patient who was admitted to the hospital from January 2012 to December 2021 and 2) The patient who was diagnosed with any infectious diseases within two days before and after the hospitalization. We used the Systematized Nomenclature of Medicine Clinical Terms code '40733004 (Disorder due to infectious disease)' 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 selected thousand admission notes randomly and manually compared notes with the HIPAA PHI list to identify the potential PHI entity in Korean clinical reports. The identified PHI entities were deidentified with the two approaches: 1) rule-based approach and 2) dictionary-based approach. Especially, since proper nouns (e.g., name, city, country, and hospital) should be considered precisely, we constructed dictionaries per each relevant PHI entity. To identify patient's names, we constructed a name dictionary containing 47,699 names made of combining last names with first names which the most preference for each year from 1940 to 2019, added with the established name dictionary. We also extracted the names of country, hospital, city and state from public open data and constructed each dictionary to identify them. Regular expression rules are also constructed to identify other PHI patterns.

3. Feature identification using topic modeling

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 can demonstrate its 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 the semantic meanings of topics.

Results

1. Data preparation

A total of 44,415 patients were identified and their 61,379 admission notes were used in the study. We eliminated the numbers and the punctuations of admission notes for pre-processing for the LDA application. Additionally, the authors decide that unit-related words (i.e., 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. The number of notes with PHI identified through the rule is as follows (name of clinician, 200; family relationship, 442; contact, 928; birthplace and residence, 240; hospital name, 21,466; hospital room number, 61,379; profession, 143)

The most of the PHI identified through the rules were de-identified, but when de-identified through dictionary-based approach, quite a few false negatives are showed due to problems of agglutinative characteristics of Korean.

3. Feature identification using topic modeling

According to the perplexity score algorithm, we found that number of 5 or 9 topics were optimal number of topics. However, the authors decided to review the only number of topics with 6 for a clear explanation of the semantic meanings, and LDA with the six topics was used in the study.

Figure 1 shows the top 11 frequencies of terms for whole documents. The number of words used in LDA was 399, and the total frequency of whole words was 2,185,836. We found "fever" has the highest frequency of whole documents (50,701/2,185,836; 2.3%), and infectious disease-related words (i.e., fever, pain, pneumonia, cough, anti, antibiotics) also showed high frequency.

Figure 2 shows the most frequently identified tokens per each topic. Sepsis-related words (i.e., shock, septic, sepsis) were clustered in topic 1, and urinary tract infection-related words (i.e., fever, uti, pyuria) were clustered in topic 2. Pediatric infection-related words (i.e., 환아로 [pediatric patient in Korean]), seizure, 해열제 [antipyretic in Korean]) were clustered in topic 3, and surgical infection-related words (i.e., insertion, dyspnea, stent) were clustered in topic 4. Respiratory infection-related words (i.e., pneumonia, cough, uri, antibiotics) were clustered in topic 5, and viral infection-related words (i.e., hepatitis, bviral, hbv) were clustered in topic 6.

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, signs and symptoms related to infectious disease in de-identified clinical notes were extracted using LDA model. This framework can be used for future studies such as data standardization of infectious disease and cohort phenotyping.

Acknowledgment

This research was supported by a grant for the project for Infectious Disease Medical Safety, funded by the Ministry of Health, Republic of Korea (grant number: HG22C0024). This research was also funded 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 (KHIDI), funded by the Ministry of Health &Welfare, Republic of Korea (grant number: HR16C0001).

References/Citations

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}	

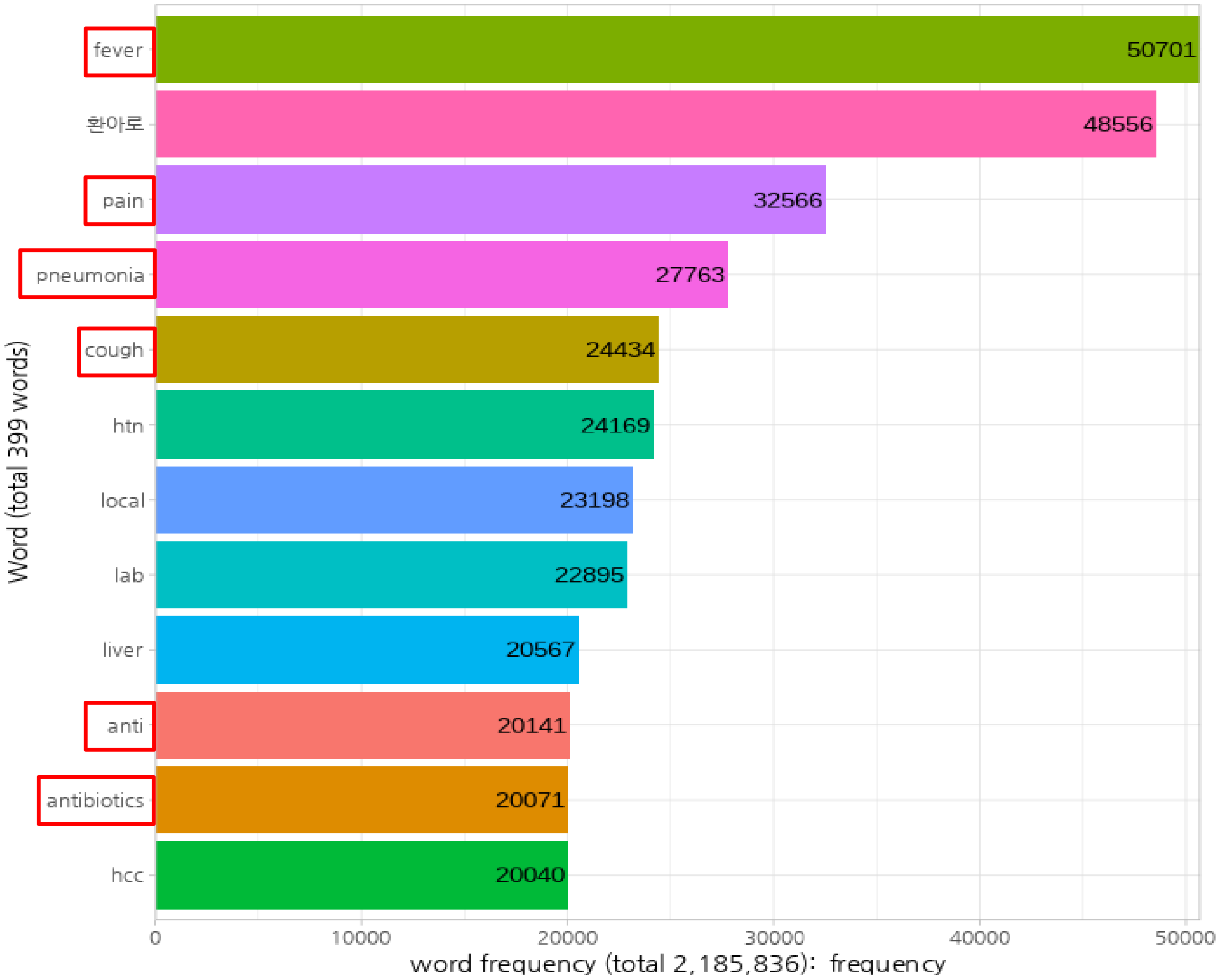

Figure 1. Word frequency plot for total documents

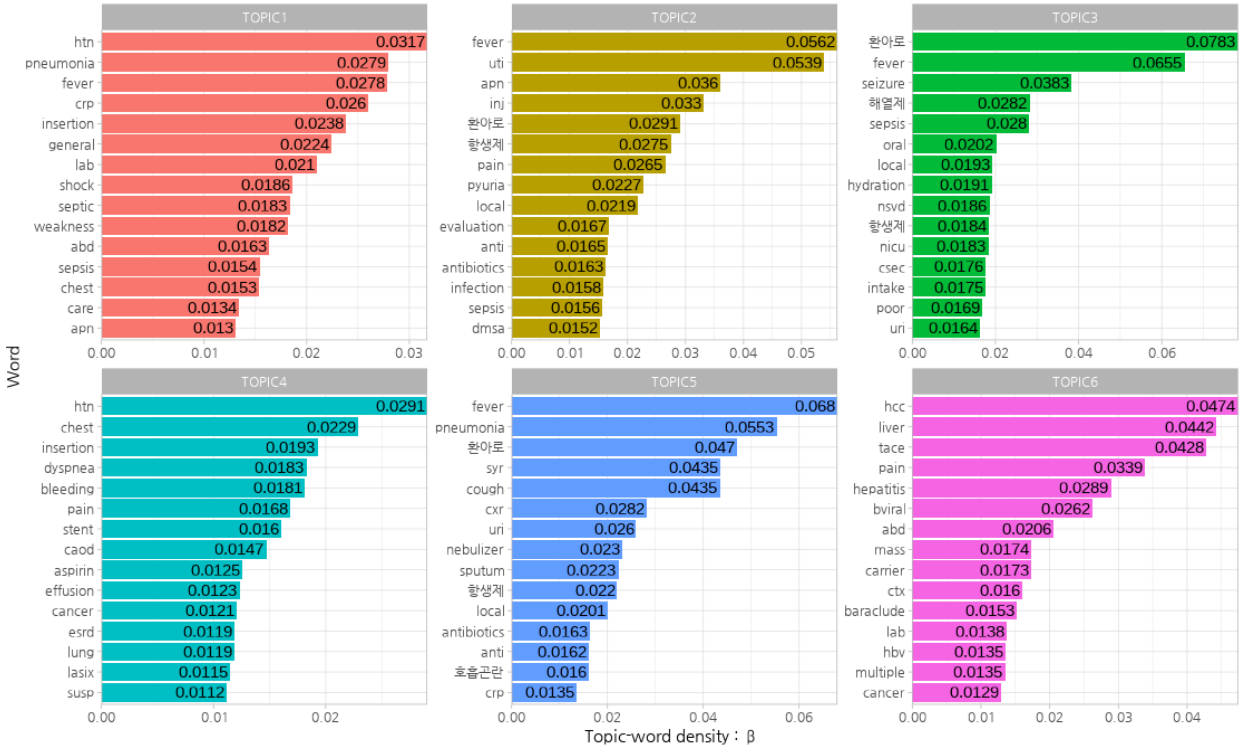

Figure 2. Word density plot for four topics