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

PRESENTER: **Junhyuk** Chang

INTRO

- Infectious disease-related information is usually recorded in the form of free-text, which needs natural language processing (NLP) to apply.
- However, most of free-text is containing protected health information (PHI) that should be de-identified.
- In this study, we applied the NLP to confirm the distribution of infection-related information after de-identifying PHI in admission notes.

METHODS

1. Data preparation

- Ajou University Medical Centre database
- Inclusion criteria

1) Admitted from Jan 2012 - Dec 2021.
2) Diagnosed with infectious disease within ±2 days from the admission date.
   - Infectious disease diagnosis
     : SNOMED code '40733004 (Disorder due to infectious disease)' and its sub-hierarchy codes

2. PHI identification and de-identification

- We compared 1,000 admission notes that were randomly selected with the HIPAA PHI list to identify the potential PHI entity.
- Two approaches to de-identify PHI entities

1) Dictionary-based approach	2) Rule-based approach
• For name, country, and hospital entities	• For other PHI patterns

3. Feature identification using topic modeling

- Tokenization
  - By unigram
  - Descriptive analyses for frequency
- Latent Dirichlet allocation (LDA)
  - Describing documents by clustering words based on the frequency
  - Perplexity score to decide an optimal n of topics

RESULTS

Extract admission notes and PHI de-identification

- We extracted patients and their admission notes.
- We identified PHI entities and their patterns.

44,415 patients → **61,379** Admission notes → **9** PHI **21** Pattern

- Constructed dictionaries (dictionary : cases)
  - Name : 47,696, Country : 241, Hospital : 45,932

(regular expression rules to de-identify showed in the abstract).

Infectious disease can be screened and detected through natural language processing after de-identifying patient health information

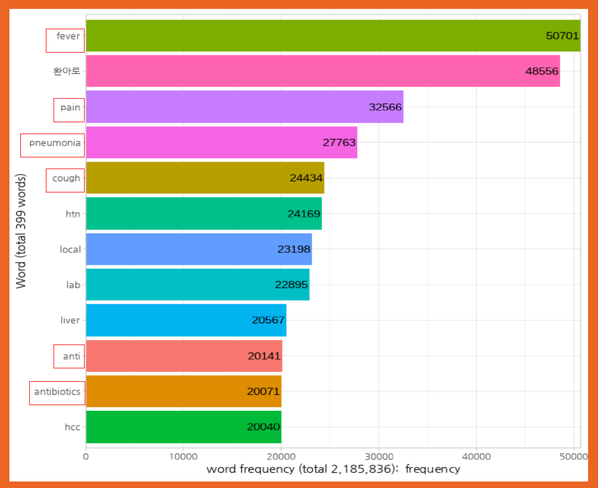

Figure 1. Word frequency plot for total documents

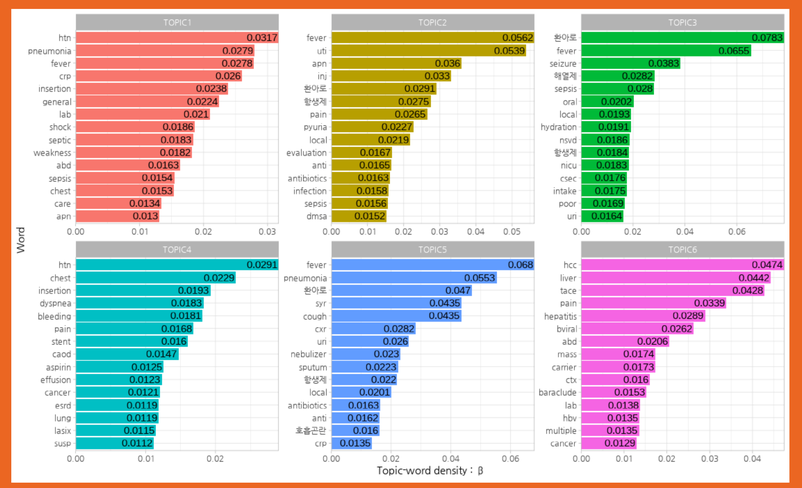

Figure 2. Word density plot for four topics.

Scan QR to **download** the **abstract** or **poster**.

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

Descriptive summary

- "**fever**" has the highest frequency (50,701/2,185,836 ; 2.3%) (Figure 3).
- **Infectious disease related words** (red box) also showed high frequency.

LDA topic modeling

- Decided optimal topic number
  - 5~9 topics were the optimal topic number according to the perplexity score
  - **6 topics** for a clear explanation of semantic meanings

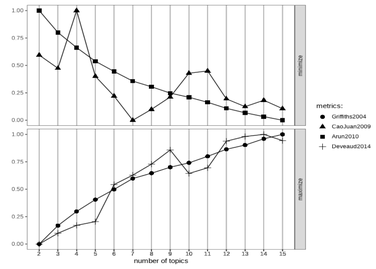

Figure 3. Perplexity scores plot

- Figure 2 shows the most frequently identified words per each topic.
- Clustered word per each topic related below.

Topic 1 Sepsis	**Topic 2** Urinary tract infection	**Topic 3** Pediatric infection
Topic 4 Surgical infection	**Topic 5** Respiratory infection	**Topic 6** Viral infection

- Relevance of clustered words per each topic (Figure 4).

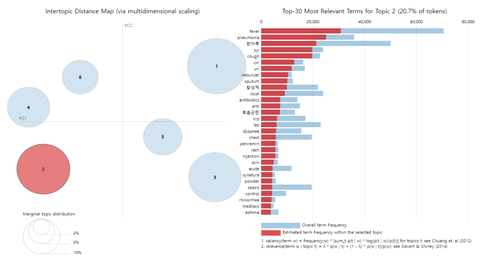

Figure 4. Topic distance map and relevant terms for the topic 2

CONCLUSION

- In this study, we extracted sign and symptoms related to infectious disease from deidentified clinical records using natural language processing technique.
- This framework can be used for future research such as data standardization of infectious disease and cohort phenotyping.

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
[2]Department of Biomedical Informatics, Ajou University School of Medicine