Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230156

Text Extraction and Standardization System Development for Pathological Records in the Korea Biobank Network

SooJeong KO[a,b] , Sunghyeon PARK[a,b] , SeolWhan OH[a], YunSeon IM[a,b], Surin JUNG[a], BoYeon CHOI[a,b], Jaeyoon Kim[a,b], Wona CHOI [b,1] and InYoung CHOI [b,2]
[a] *Department of Medical Informatics, College of Medicine, The Catholic University of Korea*
[b] *Department of Biomedicine and Health Sciences, The Catholic University of Korea*
ORCiD ID: SooJeong Ko https://orcid.org/0000-0002-6550-9188
Sunghyeon PARK https://orcid.org/0000-0002-2235-4358
SeolWhan OH https://orcid.org/0000-0002-0328-9634
YunSeon IM https://orcid.org/0000-0002-2510-1380
Surin JUNG https://orcid.org/0000-0002-3314-3185
BoYeon CHOI https://orcid.org/0000-0003-1311-2645
Jaeyoon Kim https://orcid.org/0000-0001-8847-9586
Wona CHOI https://orcid.org/ 0000-0003-0269-6374
InYoung CHOI https://orcid.org/ 0000-0002-2860-9411

Abstract. In Korea, the Korea Centers for Disease Control and Prevention operates the Korea BioBank Network (KBN). KBN has pathological records that collected in Korea and it is useful dataset for research. In this study, we established system that time efficient and reduced error by step-by-step data extraction process from KBN pathological records. We tested the extraction process by 769 lung cancer cohorts and 1292 breast cancer cohorts and accuracy is 91%. We expect this system can be used to efficiently process data from multiple institutions, including Korea BioBank Network.

Keywords. NLP, BioBank System,

1. Introduction

Pathological records are essential for many studies such as cancer diagnosis, treatment, and prognosis, but they are usually written in free text format and need to be converted to a standardized and structured format. In Korea, the Korea BioBank Network collects and distributes biobank samples and clinical data [1], but data collection methods vary and manual data entry can lead to errors. Recently, rule-based natural language processing (NLP) systems have been used to extract structured genotype information

[1] Corresponding Author: Wona Choi, E-mail choiwona@gmail.com
[2] Corresponding Author: InYoung Choi, E-mail: iychoi@catholic.ac.kr

from free-text reports [2,3]. To improve data quality, this research finds efficient ways of collecting essential data from pathology records must be found to minimize errors

2. Methods

This study presents a system that uses rule-based extraction to quickly and accurately retrieve data from pathological records texts. The system utilizes Python and its packages to create a user-friendly interface with a step-by-step process that reduces error rates. The process involves creating a sample population, validating and checking for missing data, setting a Large Frame and a Small Frame, saving inserted terms and extracted words as a dictionary, and extracting the final result from the text of the pathological records.

Figure 1. Flow chart of the extraction process

3. Results

This study tested the extraction process on 769 lung cancer and 1292 breast cancer cohorts, achieving an extraction rate of 95% and 91%, respectively. The process was also successful in standardizing the information. Additionally, the time taken for the process was 15 minutes for lung and breast cancer records, as opposed to the previous manual process which took over 1 hour. This suggests a significant reduction in the time required for information extraction.

4. Discussion

This study presents a system for automated data extraction and management from multiple institutions record. The system is designed to allow researchers with knowledge of pathology recording technology and clinical expertise to easily extract and manage data, while allowing the system to check the extraction rate and update extraction conditions for changed terms and forms. Additionally, the system can be installed in medical data management platforms for efficient and effective ETL for quality verification.

References

[1] Ko SJ, Choi W, Kim KH, Lee SJ, Min H, Oh SW, Choi IY. Common Data Model and Database System Development for the Korea Biobank Network. Applied Sciences. 2021 Jan;11(24):11825.
[2] Lee KH, Kim HJ, Kim YJ, Kim JH, Song EY. Extracting structured genotype information from free-text HLA reports using a rule-based approach. Journal of Korean Medical Science. 2020 Mar 30;35(12).
[3] Chen L, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. Journal of the American Medical Informatics Association. 2019 Nov;26(11):1218-26.