SOCRATex - a Staged Optimization system for Curation, Regularization, and Annotation of clinical Text

Jimyung Park[1], Seng Chan You, M.D., M.S.[2], Jin Roh, M.D.[3], Ph.D, Dongsu Park[2], Kwang Soo Jeong[2], Rae Woong Park, M.D., Ph.D[1,2]

[1]Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Yeongtong-gu, Suwon, South Korea; [2]Dept. of Biomedical Informatics, Ajou University School of Medicine, Yeongtong-gu, Suwon, South Korea; [3]Dept. of Pathology, Ajou University Hospital, Yeongtong-gu, Suwon, South Korea

Is this the first time you have submitted your work to be displayed at any OHDSI Symposium?

Yes __V__ No_____

Abstract

Unstructured clinical narrative reports may contain invaluable information. To extract precious information from clinical reports, free text must be converted into machine-readable data via Natural Language Processing (NLP). However, Korean clinical reports can be difficult to process for reasons including multilingual contents and the use of different formats. There is still a lack of NLP tools for Korean clinical reports. Hence, a user-friendly NLP tool which can be accommodate the diverse formats of Korean clinical reports is needed. In this research, we aimed to develop a system called SOCRATex, which is optimizing the stages of NLP, including preprocessing, exploration, and annotation of clinical narrative text. SOCRATex is developed based on OMOP-CDM (Observational Medical Outcome Partnership −Common Data Model), so it can be used in other hospitals and for different types of clinical reports. To assess the feasibility of SOCRATex, we applied it to the analysis of pathology reports on colorectal cancer at Ajou University Medical Center.

Introduction

Since unstructured clinical reports are in the form of natural language, they need to be processed and annotated using Natural Language Processing (NLP) for computational accessibility and analysis[1,2]. However, this annotation requires medical experts to review all of the reports, an extremely time-consuming process. The analysis and storage of annotated data is not always possible, due to a lack of suitable tools and databases. Our aim was to establish a usable system by optimizing preprocessing, exploration, annotation, and search engine access based on OMOP-CDM, to enable multi-center distributable research.

Methods

The pathology reports for colorectal cancer patients from 2014 to 2017 were obtained from Ajou University Medical Center. Colorectal cancer was defined using the SNOMED-CT codes, which were converted from the C18-20 codes of ICD-10 (International Classification of Disease, 10th revision). The pathology reports of other anatomical sites were excluded.

The pipeline of SOCRATex consists of four main processes: preprocessing; exploration; annotation; and searching engine (Figure 1). The algorithms used for preprocessing were aimed at the removal of punctuation, the establishment of a user-defined dictionary, and similar data cleansing. During exploration, Latent Dirichlet Allocation (LDA) can be applied to the preprocessed data for topic clustering[3].

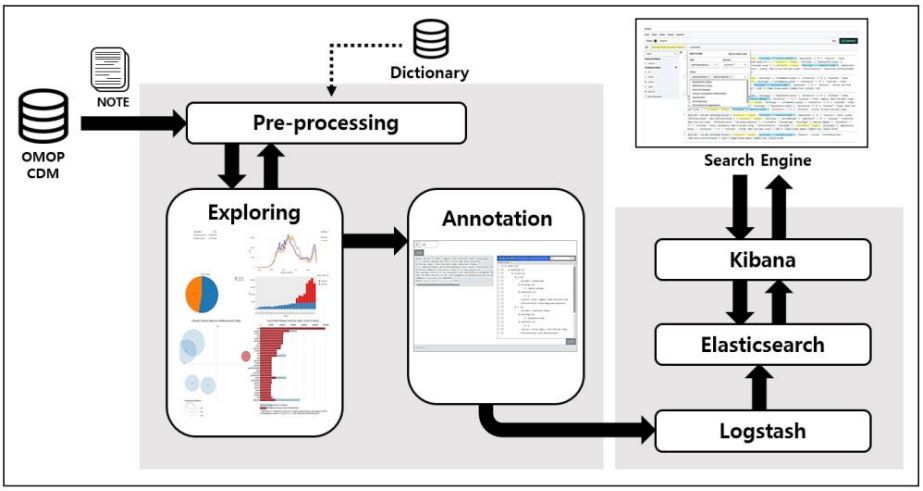

Figure 1. The architecture of the SOCRATex, consisting of preprocessing, exploration, annotation, and ELK stack (Elasticsearch, Logstash, Kibana)

After the identification of topic clusters showing sample documents, researchers can define JSON (JavaScript Object Notation) objects and templates (JSON schema). JSON data can be validated using JSON schema, which forces the use of JSON data types and lists according to the user-definition. Finally, a clinical text search engine can be constructed by sending the JSON data to Elasticsearch. Elasticsearch was developed using the open source software Apache Lucene, and offers real-time searching on the inverted indices of JSON data.

The user-friendly system was developed for researchers who are unfamiliar with NLP techniques or advanced programming skills, using *R version 3.4.4* and the interactive web application development framework *Rshiny.*

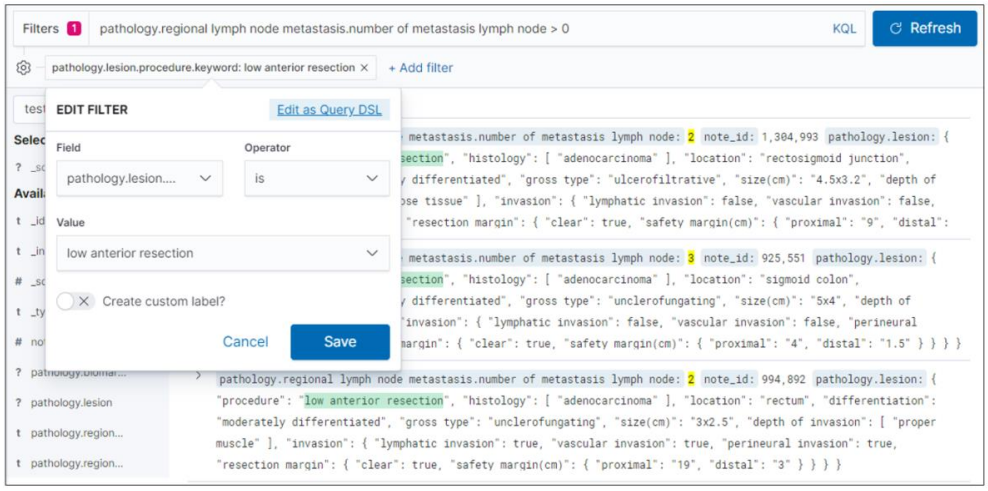

Figure 2. Sample view of Kibana, searching for pathology reports of low anterior resection with metastasis lymph nodes.

Results

Overall, 1,985 pathology reports from 1,929 patients with colorectal cancer from 2014 to 2017 were extracted. Only the pathology reports were extracted, by user selection. Abbreviations and stop words such as *a-colon*, *ti*, *dsc*, *bx,* were deleted or replaced using the dictionary and other preprocessing algorithms. LDA, a generative probabilistic model, was used to cluster text into five topics. The results of the topic clustering are listed in (Table 1). The JSON schema was defined mainly using a pathology result. The pathology result is composed of regional lymph nodes, biomarkers, and lesions consist of annotations such as histology, location, differentiation, gross type, invasion, and resection margins. The JSON data is sent to Elasticsearch to construct the clinical text search engine (Figure 2).

Conclusion

SOCRATex was used to demonstrate the process of clinical document annotation and construction of the search engine using validation with pathology reports about colorectal cancer. SOCRATex allows users to extract, process, and annotate clinical text in OMOP-CDM. To validate the usability and feasibility of the system, 1,985 pathology reports from Ajou University Hospital were analyzed. The reports were transformed into JSON format and sent to Elasticsearch. Consequently, a scalable clinical text search engine based on OMOP-CDM was created and validated using 1,985 pathology reports.

Table 1. Top most 30 terms of topics from Latent Dirichlet Allocation

Topics	Terms
Topic1	***biopsy**, all, consists, xxcm, embedded, mucosal, received, measuring, diagnosis, sections, tissue, pieces, labelled, gross, **biopsied**, **adenocarcinoma**, **cancer**, **differentiated**, colon, **moderately**, rectal, verge, rectum, anal, four, sigmoid, endoscopic, largest, five, one*
Topic2	*anal, verge, colon, one, **tubular**, **adenoma**, **low**, **grade**, **dysplasia**, biopsy, transverse, **polypectomy**, containers, each, ascending, identified, consists, two, polyp, largest, sigmoid, descending, polypoid, hyperplastic, mucosal, proximal, endoscopic, polyps, xxcm, three*
Topic3	*margin, resection, mass, **lymph**, **invasion**, **node**, **regional**, xcm, **metastasis**, distal, apart, len, **pericolic**, identified, proximal, carcinoma, circumference, **nodes**, fresh, some, free, illdefined, state, cut, instability, test, msimicrosatellite, bat, invades, iple*
Topic4	***invasion**, **adenoma**, resection, margin, submitted, consu, ation, hampe, grade, histopathologic, stained, size, **adenocarcinoma**, **dysplasia**, high, **tumor**, **tublovillous**, type, depth, low, biopsy, gross, **well**, **tubular**, labelled, **differentiated**, polypectomy, colon, endoscopic, whitish*
Topic5	***kras**, **mutation**, **analysis**, **dna**, **realtime**, clamping, **pcr**, codon, comments, **antiegfr**, rapy, msi, using, **genomic**, isolated, mediated, paraffinembedded, target, cetuximab, panitumumab, **marker**, pnamediated, materials, erlotinib, gefitinib, kinase, tyrosine, inhibitor, pna, additional*

Acknowledgement

This work was 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 : HI16C0992] and the Bio Industrial Strategic Technology Development Program (20005021) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

1. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform. 2018;77:34-49.
2. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med. 2016;66:29-39.
3. Sievert C, Shirley KE. LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 2014;pp 63–70.