SOCRATex Project

Development of a Scalable Search Engine for Clinical Narrative Text in OMOP-CDM

Jimyung Park*, Seng Chan You M.D. M.S.**, Jin Roh M.D. Ph.D†, Dongsu Park***, Kwang Soo Jeong**, Rae Woong Park M.D. Ph.D**,***

* Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Yeongtong-gu, Suwon, 16499
** Dept. of Biomedical Informatics, Ajou University School of Medicine, Yeongtong-gu, Suwon, 16499
†Dept. of Pathology, Ajou University Hospital, Yeongtong-gu, Suwon, 16499
Background

- To unlock the information in narrative text, annotation and processing using Natural Language Processing (NLP) is necessary
- Recent NLP techniques use labeled and unlabeled data resulting from NLP tools

Background

• Most of existing NLP tools are developed based on EHR system
• Since the EHR systems are diverse according to the institutions, reusing the NLP tools is challenging
Can a single process be applied to several databases?
Is it possible to build one single data standard for narrative text?
Can NLP tool generated labeled or unlabeled data be utilized for analysis?
• **SOCRATex**
  
  Staged Optimization of Curation, Regularization, Annotation of clinical Text

• A system of processing NOTE documents in OMOP-CDM
  ✓ Topic model (Latent Dirichlet Allocation)
  ✓ JSON schema and annotation
  ✓ ELK stack (Elasticsearch, Logstash, Kibana)
  ✓ Analysis using annotated JSON data

• **SOCRATex** aims to build a scalable and extensible NLP system based on OMOP-CDM
Web Application

OMOP CDM → Select note types → NOTE → Preprocess → Latent Dirichlet Allocation → User Insight and Curation → Data Template Definition → Annotation

User Interaction

User Insight and Curation

Data Template Definition

Analysis and Search Engine

Visualization Kibana → Elasticsearch Logstash → JSON → Storing

Analysis → Database → Storing Indexing
Method

- NOTE contains diverse types of documents and contents
- ATHENA provides essential concept_ids for note type classification, yet subtypes are not provided
- Reviewing the whole reports is necessary to detect subtypes
Method: LDA clustering

- To avoid reviewing all clinical reports, clustering is needed to see through into the documents
- Latent Dirichlet Allocation (LDA) can detect and classify the topics among the documents
- The topics can describe the contents of the documents
Method: JSON Schema andAnnotations

JSON schema

```json
{
    "$id": "https://example.com/person.schema.json",
    "$schema": "http://json-schema.org/draft-07/schema",
    "title": "Person",
    "type": "object",
    "properties": {
        "firstName": {
            "type": "string"
        },
        "lastName": {
            "type": "string"
        },
        "age": {
            "type": "integer",
            "minimum": 0
        }
    }
}
```

JSON data

```
{
    "firstName": "John",
    "lastName": "Doe",
    "age": 21
}
```

- With clustering result, user can get an information to define a JSON schema which contains the documents
- JSON schema defines the structure of JSON which ensures the quality of data
- After defining JSON schema, user need to annotated JSON based on the template
Method : ELK stack

- The annotated JSON can be sent to Elasticsearch
- Elasticsearch is a full text search engine with a schema-free JSON documents
- By building ELK stack with annotated JSON data, clinical text search engine can be constructed

https://github.com/elastic/elasticsearch
Method: Validation

- Ajou University Medical Center
- ICD-10th C18-20 diagnosed from 2014-2017 were included
  - Malignant neoplasm of colon, rectosigmoid junction and rectum
- 1,989 pathology reports on colorectal cancer of 1,929 patients were included
Results: LDA clustering

margin, resection, mass, lymph, invasion, node, regional, xcm, metastasis, distal, apart, len, pericolic, identified, proximal, carcinoma, circumference, nodes, fresh, some, free, illdenfined, state, cut, instability, test, msimicrostatellite, bat, invades, iple
Results: LDA clustering

- 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
### Results: LDA clustering

<table>
<thead>
<tr>
<th>Topic</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic1</strong></td>
<td>Malignant, biopsy</td>
</tr>
<tr>
<td>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</td>
<td></td>
</tr>
<tr>
<td><strong>Topic2</strong></td>
<td>Benign, biopsy</td>
</tr>
<tr>
<td>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</td>
<td></td>
</tr>
<tr>
<td><strong>Topic3</strong></td>
<td>Lymph node invasion, surgery</td>
</tr>
<tr>
<td>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</td>
<td></td>
</tr>
<tr>
<td><strong>Topic4</strong></td>
<td>Cancer, surgery</td>
</tr>
<tr>
<td>invasion, adenoma, resection, margin, submitted, consu, ation, hampe, grade, histopathologic, stained, size, adenocarcinoma, dysplasia, high, tumor, tublovillus, type, depth, low, biopsy, gross, well, tubular, labelled, differentiated, polypectomy, colon, endoscopic, whitish</td>
<td></td>
</tr>
<tr>
<td><strong>Topic5</strong></td>
<td>Gene mutation analysis</td>
</tr>
<tr>
<td>kras, mutation, analysis, dna, realtime, clamping, pcr, codon, comments, antiegfr, rapy, msi, using, genomic, isolated, mediated, paraffinembedded, target, cetuximab, panitumumab, marker, pna mediated, materials, erlotinib, gefitinib, kinase, tyrosine, inhibitor, pna, additional</td>
<td></td>
</tr>
</tbody>
</table>
Results: JSON Schema

- pathology
  - lesion
    - procedure
    - histology
    - annotation
    - location
    - differentiation
    - gross-type
    - size(cm)
    - depth-of-invasion
    - underlying-pathology
    - invasion
      - lymphatic-invasion
      - vascular-invasion
      - perineural-invasion
      - resection-margin
    - number-of-metastasis-lymph-node
      - number-of-whole-lymph-node
    - type
    - biomarker
      - result
      - method
    - regional-lymph-node-metastasis
      - clear
      - safety-margin(cm)
Results: JSON Annotation

Biopsy Result:
- A. Colon, sigmoid, 28cm from anal verge, polypectomy, Tubular adenoma with low to focal high grade dysplasia.
- B. Rectum, upper, 15cm from anal verge, endoscopic biopsy, Adenocarcinoma, well differentiated.

Gross Result: Received in for (Clinical diagnosis: R/O Rectal cancer, # A colon polyp # A)
The specimen consists of two containers each identified as A) sigmoid, 28
Each specimen consists of one, four fragments of polypoid mucosal tissue
Summary of sections: All embedded.
Block: (2)
Results: ELK stack
Results: ELK stack
Results: ELK stack

- EGFR
- kras
- nras
- CDX2
- P53
- MSI
- CK20
- MLH1
- MLH2
- BRAF
- biopsy
- polypectomy
- endoscopic biopsy
- low anterior resection
- colonoscopic biopsy
- hentman operation
- EGFR
- kras
- nras
- CDX2
- P53
- CK20
- MSI
- MLH1
- MLH2
- BRAF
- MSH6

- low anterior resection
- endoscopic biopsy
- biopsy
- polypectomy
- nras
- CDX2
- P53
- CK20
- MSI
- MLH1
- MLH2
- BRAF
- MSH6
Results: Analysis

- Using annotated JSON data, TNM stage was extracted using ‘depth of invasion’ and ‘metastasis lymph node’ keys.
- 5-year survival analysis was conducted.
- As a result, it shows similar survival rates to previously known survival rates.
- This analysis indicates that SOCRATex combined with OMOP-CDM can result in significant clinical findings.

**Numbers at Risk**

<table>
<thead>
<tr>
<th></th>
<th>Stage III</th>
<th>Stage I, II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers</td>
<td>107</td>
<td>350</td>
</tr>
<tr>
<td>Survival</td>
<td>84</td>
<td>242</td>
</tr>
<tr>
<td>Patients</td>
<td>59</td>
<td>180</td>
</tr>
<tr>
<td>Deaths</td>
<td>44</td>
<td>118</td>
</tr>
<tr>
<td>Relapse</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Censored</td>
<td>15</td>
<td>34</td>
</tr>
<tr>
<td>Events</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

5-year Survival Analysis Colorectal Cancer TNM Stage

\[P = 0.019\]
Conclusions

• SOCRATex can be applied to any type of clinical documents on OMOP-CDM
• It helps users to explore and cluster text data through a topic model and a search engine
• SOCRATex can produce labeled and unlabeled data which can be used for clinical analysis
• However, curation and annotation still needs human intervention and manual annotation
Thank you

Q&A