DEVELOPMENT OF NATURAL LANGUAGE PROCESSING (NLP) WORKBENCH WEB SERVICES

Sandy Jones
Public Health Advisor
Cancer Surveillance Branch, CDC

NAACCR 2017 Annual Conference
June 19-22, 2017
NLP Workbench Web Services
Team Members

**CDC:**
- Sandy Jones (PI)
- Joseph D. Rogers
- Wendy Blumenthal
- Renita Blake

**FDA:**
- Taxiarchis Botsis (PI)
- Mark Walderhaug
- Matthew Foster
- Kory Kreimeyer
- Abhishek Pandey
- Richard Forshee

**Northrop Grumman:**
- Steve Campbell
- Marcelo Caldas
- Sanjeev Baral
- Jon Patrick, Health Language Analytics Global

**Engility:**
- Wei Chen
- Judith Crumpler
- Leslie Eberhardt
- Wei Wang
- Guangfan (Geoffrey) Zhang
Patient-Centered Outcomes Research Trust Fund (PCORTF) Collaborative Project – CDC and FDA

• **Two Year Project** (July 2016 – September 2018)

• **Project Goals:**
  
  • Develop a Natural Language Processing (NLP) Workbench that utilizes Web Services for analyzing unstructured clinical information
  
  • Pilots for use in cancer registries and safety surveillance domains
  
  • Workbench will be used to develop NLP applications for other clinical domains
  
  • User-developed NLP applications will utilize the Web Service to deliver NLP processing to their clients
Overview of Project Activities

- Environmental Scan
- Stakeholder Engagement, Requirements Gathering, and Technical Design
- Prototype Development
- Pilot NLP Workbench Web Service
- Release NLP Workbench Web Services
Cancer Pathology and Biomarker Reporting

• About 90% of cancer cases require a pathology-confirmed diagnosis
• Pathology reports historically text-based
• Processing of narrative reports is very time-consuming
• College of American Pathologists (CAP) accreditation:
  • Require laboratories to use CAP Cancer Protocols
  • Developed electronic Cancer Checklists (eCCs) for Pathology and Biomarkers
• eMaRC Plus software does a good job of analyzing narrative reports but there is room for improvement
Challenges and Purpose

• Challenges:
  • CAP checklists aren’t required for biomarkers
  • No requirement for laboratories to store and/or transmit cancer data in discrete data elements
  • Inconsistencies across laboratories in terminologies, test names and data included in the biomarker reports
  • Inconsistencies in report organization and reporting in HL7 messages

• Purpose:
  • Improve extraction and auto-coding through use of machine learning techniques
  • Utilize shared model that can be expanded based on user input
# eMaRC Plus Processing of Narrative Pathology Reports

<table>
<thead>
<tr>
<th>Path Ordering</th>
<th>KSC/Renner, Robert Path Ordering Fac.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facility Name</td>
<td>41</td>
</tr>
<tr>
<td>Address No &amp; St</td>
<td>Holcombe</td>
</tr>
</tbody>
</table>

**Path Final Diagnosis**

- Diagnosis: A left prostate needle core biopsy - Prostatic intraepithelial neoplasia - Right - Radial needle core biopsy - Prostatic intraepithelial neoplasia, high-grade. COMMENT: A PINP immunostain is performed on sections of blocks A1-3 and B1-3. The findings support the morphologic interpretation. Clinical correlation and follow-up is recommended. Code 88185/88280.

**Histologic Type ICD-O-3**

- Grade 3 - Carcinoma in situ

**Grade Path Value**

- No Two- or Three- or Four-System Grade is available; unknown

**Grade Path System**

- Not a Two-, Three- or Four-System grade; unknown
Development of Training Dataset

• Collect data from at least four national laboratories for the following histopathology primary cancer sites (including multiple states):
  • Breast
  • Lung
  • Prostate
  • Colorectal

• 125 cases per cancer site from each laboratory for a total of at least 2,000 cases

• Double annotation will be completed by certified tumor registrars with a master reviewer
Cancer Pathology Use Cases

• Identify case reportability
  • Pre- and post-negation

• Extract cancer information
  • Histology
  • Primary site
  • Behavior
  • Laterality
  • Grade

• Code cancer data items to nationally adopted coding systems (ICD-O-3)
NLP Use Cases for Cancer Domain

1. **Cancer Registry eMaRC Plus System**
   - Transmit unstructured text
   - HL7 2.5.1 ORU message

2. **Narrative Pathology Reports**
   - HL7 2.5.1 ORU message

3. **Laboratory Information System**
   - Transmit unstructured text

4. **NLP Web Service**
   - Return Reportability Determination
   - Return Structured Data

5. **Repository of NLP Tools**
   - Reportability (ICD-10-CM)

6. **Structured Data**: Primary Site, Histology, Laterality, Behavior, Grade (ICD-O-3, etc.)

The diagram illustrates the integration of unstructured text from various sources through HL7 2.5.1 ORU messages to NLP Web Service, which then returns structured data and reportability determinations.
Introduction: Safety Surveillance

Medical product administration may be associated with the onset of **Adverse Events**.

Adverse Event monitoring is:

- easier in the pre-market phase with clinical trial data
- challenging in the post-market phase with high volumes of data:
  - Submitted to the FDA Spontaneous Reporting Systems (Passive Surveillance)
  - Evaluated in the FDA Sentinel Initiative (Active Surveillance)
Spontaneous Reporting Systems

FDA monitors post-market reports for medical products submitted annually to Spontaneous Reporting Systems such as:

- **VAERS**: Vaccine Adverse Event Reporting System (for vaccines)
  ~54,000 reports (for vaccines) in 2016

- **FAERS**: FDA Adverse Event Reporting System (for drugs and biologics)
  ~1,700,000 reports (for drugs and biologics) in 2016
NLP for Safety Surveillance
Selected Use Cases for the Pilot

**Use Case 1**: Extraction of key clinical information from safety reports

**Use Case 2**: Extraction of temporal information and identification of temporal relationships in safety reports

**Use Case 3**: Summarization of Adverse Event information included in the safety narratives
Use Cases 1 and 2: Clinical and Temporal Information Extraction

Relative Time
Patient received Smallpox vaccination on 4/21/2006 in left deltoid. **12 days** after vaccination he developed increased left arm pain and pleuritic substernal chest pain. 5/11/06 transferred to hospital with chest pain, right arm pain. Final dx of acute myopericarditis, serum reaction, allergic reaction, anemia, abnormal reaction to vaccine. Medical records from previous hospitalization obtained on 5/14/06 showed PMHx of Stevens-Johnson syndrome; family hx reveals patient's father had myocardial infarction.
Use Cases 1 and 2 (Extension): Visualize Clinical Information and Temporal Relationships
Use Case 3: Report Summarization

Utilize and combine:

- the structured data fields (age, sex, product names)
  with
- the clinical and temporal information from the safety narratives

To create two types of summaries per safety report:

- a brief textual summary
  and
- a structured tabular summary
Use Case 3: Build a Brief Textual and a Structured Summary

**Brief Textual Summary**
A 75 year-old male with medical history of skin cancer, cerebral infarction and prostatic hypertrophy was treated with thymoglobulin for aplastic anemia. The patient was diagnosed with severe aplastic anemia on 2011-06-18; died on 2011-06-25.

**Structured Tabular Summary**

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Age</th>
<th>Sex</th>
<th>Products</th>
<th>Calculated Onset</th>
<th>Primary Diagnosis</th>
<th>Secondary Diagnosis</th>
<th>Medical History</th>
<th>Concomitant Medications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>75</td>
<td>M</td>
<td>THYMoglobulin</td>
<td>3 days</td>
<td>severe aplastic anemia</td>
<td>pyrexia and...</td>
<td>Skin cancer, cerebral...</td>
<td>Ceftazidime...</td>
</tr>
</tbody>
</table>
Environmental Scan in a Nutshell

• Objective:
  • To identify NLP open source solutions (e.g. algorithms, methods, techniques, software, and large-scale systems) that may support certain functionalities of the NLP Workbench Web Service

• Met objective by conducting:
  • Literature Review that included the retrieval and review of the recent and relevant publications from multiple databases
  • Multi-Channel Review that included the systematic evaluation of other sources
  • Data Merger of the findings from the two reviews followed by filtering and evaluation
Overview of the Environmental Scan Process
Summary of Main Findings

• Significant development activity around the GATE and UIMA frameworks

• cTAKES, GATE, MedEx, MetaMap, and Stanford CoreNLP were widely-used systems and reported in all sources of our environmental scan

• Some popular components performing advanced functionalities, such as time extraction and co-reference resolution

• There were single tools found for more specialized tasks, such as abbreviation normalization

• A great number of applications and components supported named entity recognition

• Promising, however not open-source solutions, did not pass the availability criterion and were therefore excluded from the final list
Architectural Design
NLP Workbench Web Service Requirements

• Build on existing efforts
• Be modular, freely available and open-source
• Create and share NLP solutions and web services
• Satisfy multiple use cases and applications for various clinical subdomains
• Encode clinical data to multiple terminologies (ICD, MedDRA, etc.)
• Support structuring information according to certain standards, such as the Common Data Model
NLP Workbench Web Service
Dos and Don’ts

• Will include processes with demonstrated efficiency - *is more than a collection of general NLP components and workflows*

• Will cover certain needs - *cannot address all problems*

• Will describe the process for the generation of annotated datasets - *will not necessarily generate multiple new corpora*

• Intend to incorporate only open-source solutions equipped to support the project objectives - *will not endorse ANY existing solution*
NLP Workbench Web Service Architecture

eMaRC: electronic Mapping, Reporting, and Coding; ETHER: Event-based Text-mining of Health Electronic Records system; SC: Service Consumer; SP: Service Provider; CDM: Common Data Model
Statistical NLP (sNLP) - Machine Learning Processes

1. Define Use Case
2. Collect Data
3. Prepare, Explore & Annotate Data
4. Create Pipeline (Extract Features)
5. Develop Model
6. Train Model
7. Publish Model (Web Service)
8. Analyze & Evaluate
9. Monitor

- WB - Workbench Functionality
NLP Stakeholder Meeting

• Held on April 26, 2017 with 70+ participants
• Inform community about project plans and activities
• Gather input from stakeholders on:
  • Challenges to be aware of
  • Use cases for future inclusion
  • User requirements for workbench and web services
  • Sharing of research models for use by others
• Group will meet on a quarterly basis to provide updates and gather input
• Group email box: NLPWorkbench@cdc.gov
Q&As from NLP Stakeholders

• Will the workbench be deployed as a service for everyone to use or can it be deployed at the local level to protect personal health information (PHI)?
  • The workbench will provide two deployment options including unsecured services to process publicly available data and allow the services to be run locally to protect confidential data.

• Are you planning to have only one temporal analyzer (for example) or will you allow the user to choose from multiple analyzers?
  • We will include multiple analyzers for the user to select the preferred tool for their use case.

• Will the workbench allow the user to plug in particular lexicons, ontologies, etc.?
  • Yes.

• Will the workbench have a plugin style of architecture or be easily extendable?
  • Yes, the workbench will be easily expandable and detailed instructions will be provided.

• Will there be opportunities for universities to donate machine learning based systems?
  • Yes, instructions on how contributions can be shared will be provided.

• Will this fit into the regular cancer registrar workflow or is it more for model testing?
  • The plan is for this to be integrated into eMaRC Plus to process cancer pathology and biomarker reports. It will also be made available to any other software that need to process text-based health data.
Conclusions

• Both agencies have previously worked and developed systems for the processing of cancer and safety free-text data

• Continuously examine existing solutions and consider their inclusion in the Workbench, for example:
  • National Cancer Institute/Department of Energy
  • National Science Foundation
  • National Library of Medicine

• Leverage previous work by turning research into language engineering

• Initially focus on specific use cases for two domains however aim at wider coverage

• Achieve the maximum possible performance for the selected use cases

• Align with various stakeholders’ visions and different user roles
Go to the official source of cancer prevention information: www.cdc.gov/cancer.

Speaker Contact Information:
Sandy Jones, sft1@cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.