Development of a Natural Language Processing (NLP) Web Service for Structuring and Standardizing Unstructured Clinical Information

Sandy Jones, Public Health Advisor
Cancer Surveillance Branch
Division of Cancer Prevention and Control
Centers for Disease Control and Prevention (CDC)
Good Morning.
Background and Purpose

- Meaningful Use and development of Electronic Health Record (EHR) Systems ~ more standard clinical data
- U.S. Laboratories still produce narrative reports
- Food and Drug Administration (FDA) receives narrative data for vaccine, blood products, and drugs as part of their Spontaneous Reporting Systems
- NLP Web Service needed to convert narrative data into standard coded data
PCORTF Collaborative Project – CDC and FDA

• Two Year Project
• Project Goals:
  – Develop a Natural Language Processing (NLP) Web Service for structuring and standardizing unstructured clinical information
  – Pilot for use in cancer registries and safety surveillance domains
  – Will be expandable to meet other domain requirements
Stakeholders

- Centers for Disease Control and Prevention (CDC)
- Food and Drug Administration
- Hospitals, Cancer Centers, and Universities
- Local, State, and Regional Public Health Agencies
- Laboratories
- Department of Energy
- National Cancer Institute
- National Library of Medicine
- PCOR Researchers
Overview of Project Activities

1. Literature Review and Evaluation
2. Engage Stakeholders, Requirement Gathering, and Technical Design
3. Prototype Development
4. Pilot NLP Web Service
5. NLP Web Service Published
Potential Challenges

- Dependencies on open source software
- Availability of large test datasets
- Limitations of open source NLP software and tools
- Inclusion of stakeholders from different domains
- Prioritization of requirements
Conceptual NLP Design for Cancer Pathology

- Narrative Pathology Reports
- HL7 2.5.1 ORU message
- Cancer Registry eMaRC Plus Database
- NLP Web Service
- Repository of NLP Tools, Methods and Algorithms
- Structured Data: Primary Site, Histology, Laterality, Behavior, Grade (ICD-O-3, etc.)
- Reportability (ICD-10-CM)

Transmits unstructured text
Returns Structured Data
Acknowledgement of Co-Authors

**FDA**
- Taxiarchis Botsis

**CDC**
- Wendy Blumenthal
- David Butterworth
- Joseph D. Rogers
Questions?