Natural Language Processing to Support Cancer Registries and Cancer Surveillance

Paul Fearn, PhD, MBA
Fred Hutchinson Cancer Research Center
NCI Surveillance Research Program

pfearn@fredhutch.org, paul.fearn@nih.gov

NAACCR 2016 Annual Conference, June 15th, St. Louis, MO
Objectives of this Talk

1. Introduce the NAACCR community to natural language processing (NLP) technology
2. Provide an overview of NLP tools and methods that have been applied to cancer registries and cancer surveillance
3. Provide a vision and framework to better understand opportunities for informatics, NLP and machine learning to support cancer registration and surveillance
Question 1: How familiar are you with NLP and how it might apply to cancer registries?

1. Is NLP a type of sandwich?
2. I have some idea of what NLP is but am not sure how it could apply to my work?
3. I have at least some knowledge or experience with NLP and how it might apply to registries
4. I have considerable knowledge or expertise about NLP and experience applying it in registries
What is NLP?

• Computer text processing algorithms/models, leveraging linguistic features
  – Related to search (information retrieval, finding and highlighting key words in documents)
  – Includes speech (e.g., Siri), but that’s not our focus
  – Uses algorithms/models similar to computational biology methods (e.g., models to recognize patterns in strings, predicting labels/results using features in text)

• Methods for developing NLP algorithms
  – Rules-based
  – Statistical models
  – Machine learning
  – Hybrid models
NLP and Text Mining Methods

- Information Retrieval (e.g., Google, Bing, DuckDuckGo)
- Document Classification (e.g., MeSH, document types)
- Natural language processing (NLP)
  - Usually not optical character recognition (OCR)
  - Text or speech, but we are focusing on text documents
  - Tokenizing words in narrative text (e.g., unigrams)
  - Named entity recognition (e.g., key words, ontologies)
  - Normalizing word formats (e.g., tense, caps, abbreviation)
  - Tagging parts-of-speech (POS) of words
  - Segmenting phrases or sentences in narrative text
  - Shallow or deep parsing of phrases or sentences
  - Identifying document sections (e.g., addendum)
EGFR MUTATION MOLECULAR (PCR) STUDY:
Testing performed by: **PLACE, WA **ZIP-CODE, report
***PATH-NUMBER[2].
Interpreted by: **NAME[TTT SSS], M.D., issued: **DATE [Redacted].

Result: Subcarinal Fine Needle Aspirate: NEGATIVE for an EGFR mutation by real time PCR.

Comment: In this case no epidermal growth factor receptor (EGFR) mutations were identified by Qiagen/Dxs real time PCR. This real time PCR assay has been validated to identify the most common EGFR tyrosine kinase domain mutations that involve exons 18-21. This test is validated for non-small cell lung carcinoma. The clinical significance and use of this assay in other tumor types is not well characterized. Note that this assay is not intended to diagnose any particular type of cancer, but as an aid to clinicians, is intended to be used as an adjunct to other prognostic factors to select eligible patients for therapy. (**NAME[RRR: QQQ] et al., NEJM 2004; **ID-NUM, Mok et al., NEJM 2009; **ID-NUM, Pao et al., JClinOnc 2005; 23: 2556-68).

ALK FISH analysis has been initiated as requested, and the results will be reported separately.

Addendum #1 performed by **NAME[PPP M. OOO], M.D. Electronically signed **DATE [Redacted]
The simplest--but least robust--way of classifying a document or extracting data elements from text is to write rules (if/then statements).

**EGFR Positive**

- If “EGFR” appears within $x_1$ distance of the word “positive” and there is no other test/gene name or a new sentence between them.
- Then call it Positive.

**Training Data**

**Raw Text (X)**

PLEURAL BIOPSY:

1. POSITIVE FOR AN EXON 19 DELETION EGFR MUTATION BY REAL-TIME PCR.
2. NEGATIVE FOR AN ALK REARRANGEMENT BY FISH.
How to build more robust statistical / machine learning algorithms or models,

Vector of independent feature (X) variables --- starting with raw text as “bag of words”

Dependent (Y) variables either as binary labels (e.g., EGFR+/-) or multiple data elements
Features (X) Tokenized, Lower-Case, with Common but Uninformative Words Removed

Vector of independent feature (X) variables with less noise than in the original clinical document

Dependent (Y) variables either as binary labels (e.g., EGFR+-) or multiple data elements
Features (X) Filtered by Part of Speech

Vector of independent feature (X) variables with more noise removed

Dependent (Y) variables either as binary labels (e.g., EGFR+/-) or multiple data elements
Using vector of unigram (word) features (X) to classify report as **lung vs breast**

- Just using the words themselves as features (often called “bag of words”)

- **Invasive ductal carcinoma of the right breast**
- **Adenocarcinoma of the lower right lobe of the lung**

Slide adapted from presentation by Emily Silgard
Using vector of unigram (word) features (X) to classify a report as *sarcoma vs not sarcoma*

- Just using the words themselves as features (often called “bag of words”)

  - no evidence of sarcoma, mild necrosis
  - clear evidence of sarcoma, no necrosis

Slide adapted from presentation by Emily Silgard
Using bigram and trigram features (X) to classify a report as *sarcoma vs not sarcoma*

- What about including a little more context in feature vector (X)? (bigrams, trigrams, skip grams, etc.)

---

Slide adapted from presentation by Emily Silgard
EGFR MUTATION MOLECULAR (PCR) STUDY:
Testing performed by: **PLACE, WA** **ZIP-CODE, report
***PATH-NUMBER[2].
Interpreted by: **NAME[TTT SSS], M.D., issued: **DATE[Aug 17 2012].

Result: Subcarinal Fine Needle Aspirate: NEGATIVE for an EGFR mutation
by real time PCR.

Comment: In this case no epidermal growth factor receptor (EGFR) mutations were identified by Qiagen/DxS real time PCR. This real time PCR assay has been validated to identify the most common EGFR tyrosine kinase domain mutations that involve exons 18-21. This test is validated for non-small cell lung carcinoma. The clinical significance and use of this assay in other tumor types is not well characterized. Note that this assay is not intended to diagnose any particular type of cancer, but as an aid to clinicians, is intended to be used as an adjunct to other prognostic factors to select eligible patients for therapy. (**NAME[RRRQQQ] et al., NEJM 2004; **ID-NUM, Mok et al., NEJM 2009; **ID-NUM, Pao et al., JClinOnc 2005; 23: 2556-68).

ALK FISH analysis has been initiated as requested, and the results will be reported separately.

Addendum #1 performed by **NAME[PPP M. OOO], M.D. Electronically signed **DATE[Aug 17 2012]
Linguistic Features (X) - Shallow Parsing

• Shallow parsing:
  “no EFGR mutations were identified by real-time PCR”
  “no EGFR mutations” “were identified by real-time PCR”

Identify basic structures
NP-[no EFGR mutations] VP-[were identified...]
Linguistic Features (X) - Deep Parsing

John hit the ball.

Grammar Rules

\[ S \rightarrow NP \ VP \]
\[ VP \rightarrow V \ NP \]
\[ NP \rightarrow Det \ N \]

https://class.coursera.org/nlp
Annotation (Training Data) for Document Classification* Algorithms

• What it looks like, for each document
  – Source document (or a unique id /reference #) for the source document)
  – One output value (Y)* for the given field, e.g. one NAACCR abstract-based variable and a single pathology report
  – Features (X) derived by NLP engineer

• Example
  – Labeling radiology reports as either containing evidence of recurrence or not (binary outcome)
Document Classification

• What it’s good for?
  – Data elements, document labels, or classifications that have only one valid value for each document (e.g. “This is a lung cancer pathology report.”)

• Used for interests, authorship, plagiarism?
  – http://dejavu.vbi.vt.edu/dejavu/
  – http://www.biosemantics.org/jane/
  – Spam detection
Example of Document Classification:
What is the subject of MEDLINE article?

MEDLINE Abstract in PubMed

MeSH Subject Heading Ontology

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Adapted from presentation by Meliha Yetisgen-Yildiz, PhD
Document Level Annotation with Text Documentation (Anchors)

• What it’s good for
  – Giving slightly more information about why a certain document level value (Y) was chosen (e.g., “These histologies and sites mentioned in the report are the reason that I would classify it as a lung cancer pathology report.”)

• What it looks like
  – In addition to source document and output values (Y), there would be anchors (copied “text documentation” or annotator explanations) providing evidence for the output values, e.g. a NAACCR abstract based variable (Y), text documentation in the abstract (possible X), and one or more corresponding pathology reports (source documents)

• Example
  – Labeling radiology reports as either containing evidence of recurrence or not, adding a notes field (text documentation) to specify text that explains the decision for the label (Y)
Token Level Annotation

• What it’s good for
  – All of the previous cases, as well as for extracting values for **multiple data elements or labels (Y)** from the same source document (e.g., “These spans of text are associated with the histologies of each of the specimens listed.”)

• What it looks like
  – In addition to the source document, output values (Y), there is a precise anchor (markup in the source document itself or reference to character offsets) to the exact evidence (X) for each of the field values (Y) extracted

• Example
  – Use an annotation tool (e.g., Brat, LabKey) to highlight text in the source document (X) that links directly to each coded label or value extracted (Y)
** DEMOGRAPHICS DRAWN FROM PATHOLOGY REPORT **

<table>
<thead>
<tr>
<th>FIELD</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PATIENT</td>
<td>HUTZ, LIONEL</td>
</tr>
<tr>
<td>MRN</td>
<td>U99999995 (Springfield MC)</td>
</tr>
<tr>
<td>DOB</td>
<td>Mar 3 1948</td>
</tr>
<tr>
<td>SEX</td>
<td>M</td>
</tr>
<tr>
<td>CASE</td>
<td>SU-12-99995</td>
</tr>
<tr>
<td>COLLECTED</td>
<td>Mar 2 2015</td>
</tr>
<tr>
<td>RECEIVED</td>
<td>Mar 22 2015</td>
</tr>
</tbody>
</table>

**FINAL DIAGNOSIS:**

- Liver, core needle biopsy: Poorly differentiated non-small cell carcinoma with focal squamous features consistent with metastasis from a lung primary; see comment.

**COMMENT:**

- The carcinoma is composed of nodules of cells with increased nuclear:cytoplasmic ratios, numerous mitotic figures, and patchy necrosis. No malignant glandular formation is identified. Clinical/radiologic evaluations are the best methods of determining primary site.

- The background hepatic parenchyma contains nonspecific mild portal and lobular inflammation and focal bile ductular reaction, most consistent with mass effect, and mild macrovesicular steatosis (5%).

**CLINICAL DATA:**

- 71-year-old male with liver lesions ?metastasis, history of interstitial lung disease, sarcoid, and pulmonary fibrosis. Per Mindscape, status post left lung transplant with endobronchial mass in right lower lobe of native lung.

**GROSS DESCRIPTION:**

- Received in formalin labeled "Hutz, Lionel four cores of tissue from lesion in liver" are multiple tan-red cores and core-like fragments of soft tissue ranging from 0.1 x 0.1 cm to 0.7 x 0.1 cm. The specimen is wrapped, touched with hematoxylin, and entirely submitted in cassette A1. (RS/clk)

Kearney Zzyzwicz MD  
GI & Hepatic Fellow  
3/26/2015  
Ned Flanders MD  
Pathologist  
Electronically signed 03/6/2015

In compliance with CMS regulations, the pathologist’s signature on this report indicates that the case has been personally reviewed, and the diagnosis made or confirmed by, the Attending Pathologist. Microscopic examination was used to arrive at the diagnosis unless indicated otherwise.
1. Clinical Documents (e.g., E-Path, radiology reports) → A → 2. IMS Databases → A → 3. Text De-Identification Tool (e.g., BoB) → B → 4. VTR

A. identified documents → B. de-identified documents → C. identified documents with markup of features → D. training/validation data → E. algorithms → F. clinical data elements → G. documents with feature vectors, clinical data elements, and links between features and data elements

6. Annotation Pipeline and Task Management (LabKey)

7. Automated Annotation (Any NLP/Machine Learning/De-ID Tool) → 8. Annotation by People (LabKey) → 9. Annotation Review by People (LabKey)
NLP Engines – Collections of Algorithms to Compute Labels/Elements (Y) Given Documents/Features (X)

Input/Arguments → NLP Engine → Output/Results

Pathology
- Pathology Report Parser
- General Pathology Modules & Resources
- Clinic Notes
- Cytogenetics
- Radiology
- Surgery

Lung
- Lung Specific Pathology Modules & Resources
- Breast
- Colorectal
- Brain
- Prostate
- Sarcoma
The cycle of advancing and improving automation

- Use automation to speed up and improve reliability of manual data processing
- Review/correct results
- Apply/calibrate NLP algorithms
- Create training/validation data
- Use manual workflow to improve automation of data processing

Slide adapted from presentation by Emily Silgard
### Evaluation of accuracy and completeness in NLP algorithms

<table>
<thead>
<tr>
<th></th>
<th>Documented</th>
<th>Not documented</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Retrieved/extracted</strong></td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td><strong>Not retrieved/extracted</strong></td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

**Precision/PPV** is a measure of accuracy, and **recall/sensitivity** is a measure of completeness. Also use **F-measure**, a weighted or balanced harmonic mean of precision and recall.

---

Vision and framework to better understand opportunities for NLP and machine learning to support cancer registration and surveillance

• Data acquisition and information processing relies on expert cancer registrars
  – Highly trained, experienced, certified staff
  – Highly engineered and managed workflows

• But, registries feeling the pain of increased scope and demand for detailed clinical data, more timely and cost efficient processing, all while keeping the data quality high
Clinical data in relation to cancer as a chronic, progressive disease

Pre-diagnosis

Diagnosed

Primary treatment (e.g. surgery)

Response to primary treatment

Recurrence

Secondary treatments (e.g. chemotherapy)

Response to secondary treatments

Recurrence
Amount of Desired Clinical Data Elements that Come from “Unstructured” Documents

• We estimated that at least 65% of desired cancer clinical data elements come from unstructured text.

• Similar analyses in many other domains found the number of data elements from unstructured sources to be anywhere from 45% to 80%.

• Efforts to advance templated clinical notes are underway, changes in workflow to adopt templates have been slow.

The Four V’s of Big Data

Volume

Variety

Veracity

Velocity

Security

What is NLP vs. what you do today?

• Today
  – Keyword search for case finding and assessing reportability
  – Text documentation during abstraction
  – Abstraction and coding rules
  – Edit checks, because neither sources nor abstractors are perfect
  – Ongoing training and quality improvement
  – How you train, measure, and improve staff

• Future
  – Computer recognizing and documenting more features (X) in text
  – Hybrid of rules, statistical models, and machine learning models to predict outputs (Y) for casefinding, reportability, and extracting/coding data elements from text
  – Optimize human/computer interactions
  – Meet human accuracy/completeness (precision/recall), exceed human reliability and ability to search for “needles in haystacks” in increasingly high volume of clinical documents
  – Doing more with limited resources
Conclusions

• Challenges of scale, variability, and security to apply NLP to registries.
  – Great opportunity for NLP and machine learning research
  – Dearth of published applied NLP research for registries
  – Registries should learn about NLP and machine learning methods and how they might be applied in registry operations
• Need literature review of adjacent areas of clinical NLP and ML in cancer domain that may be applicable to registries
• How to apply locally developed NLP tools and methods to registries, at enterprise and multi-institutional scale
• Challenges in development of training and validation data
  – Targeted annotation of clinical documents
  – Capturing training and validation data during workflow
Question 2: Where would it be most useful to have computer algorithms and NLP to assist with your work?

1. Case-finding
2. Determining whether cases are reportable or not
3. Extracting currently collected clinical data elements from text documents (e.g., pathology reports)
4. Extracting new clinical data elements from text documents (e.g., biomarkers from pathology reports, recurrence/progression from pathology or radiology reports)
Acknowledgments

• Fred Hutch (Emily Silgard)
• LabKey Software
• NCI Surveillance Research Program

Contact Information:
 pfearn@fredhutch.org, paul.fearn@nih.gov