Natural Language Processing in the Division of Vital Statistics

APRIL 16, 2019

PATRICK DRAKE

NATIONAL CENTER FOR HEALTH STATISTICS
Presentation Objectives

1. Highlight selected Natural Language Processing (NLP) approaches being used by the National Center for Health Statistics (NCHS)

2. Discuss NLP projects within the Division of Vital Statistics with real-world examples
   - Finding and classifying drug related infant deaths
   - Automating the classification of cause of fetal deaths
Use cases for Natural Language Processing

1. Searching for a topic through large volumes of text

2. Cleaning and homogenizing language prior to analysis
   a) Stemming and lemmatization
   b) Abbreviation handling
   c) Correcting misspellings

3. Learning about the language being used
   a) Finding a word’s synonyms, antonyms
   b) Are deaths from novel drugs appearing in our data? (Both legal and illicit drug use is of interest here)

4. Assigning cause of death to death certificates
Pattern Matching

Terms
- Regular expressions – a sequence of characters that define a search pattern
- Tokenization - preprocessing step where text is segmented into plausible units (i.e., tokens).
- Token – can be words, acronyms, abbreviations, numbers, punctuation symbols, etc.

Challenges
Abbreviations (MD = doctor, state?), apostrophes, hypens, varying formats (e.g., acetyl-fentanyl, acetyl fentanyl, acetylfentanyl), varying boundary demarcations (e.g., The oil prices fell in the U.S.).
Replacing abbreviations in text with their meaning during data cleaning and processing can improve the performance of any text analysis or algorithm.

Medical data (death certificates, and health records) in particular contains a wide variety of abbreviations:

- Diseases and syndromes (e.g. CM = Chiari malformation, dm = diabetes mellitus, . . .)
- Short hand (e.g. fx = fracture, hb = hemoglobin, . . .)
## Abbreviation handling:

<table>
<thead>
<tr>
<th>Input text</th>
<th>Pattern</th>
<th>Quality</th>
<th>Output text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gestational dm</td>
<td>“dm”</td>
<td>Worst</td>
<td>Gestational diabetes mellitus</td>
</tr>
<tr>
<td>NAS</td>
<td>“[Nn][Aa][Ss]”</td>
<td>Better</td>
<td>Neonatal abstinence syndrome</td>
</tr>
<tr>
<td>h.s.v</td>
<td>“[Hh][- .]<em>[Ss][- .]</em>[Vv]”</td>
<td>Best</td>
<td>Herpes simplex virus</td>
</tr>
</tbody>
</table>

**Special Characters for use with Regular Expressions and their meaning: (in R)**

**Quantifiers:**
- `*` match at least 0 times
- `+` match at least 1 times
- `?` Match at most 1 time
- `{n}` match n times

**Specifying position:**
- `^` match at start of string
- `$` match at end of string
- `\b` “word boundary” matches at end/beginning of word
Spell-checking literal text fields

Spelling errors are common in text describing health conditions, medical jargon, and descriptions of deaths.

Without handling errors in some way, a model will treat different spellings of a word as entirely unrelated.

Example:
Does “gestation iabetes and placental abruption” equal “gestational diabetes and placental abruption”? 
Spell-checking literal text fields

A good spell checker has three main components:

1. Dictionary
2. A method of measuring the “distance” between two strings
3. Language model or decision rules about which word from the dictionary was misspelled in the text

NOTE: The quality of all three parts corresponds to the overall quality of the spell checker. A bad dictionary, poor choice of distance metric, or an improper language model will cause poor results even if the other elements are well implemented.
Spell-checking literal text fields

“This sentence contains a misspleling.”

<table>
<thead>
<tr>
<th>Words/tokens</th>
<th>In dictionary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>This</td>
<td>True</td>
</tr>
<tr>
<td>sentence</td>
<td>True</td>
</tr>
<tr>
<td>contains</td>
<td>True</td>
</tr>
<tr>
<td>a</td>
<td>True</td>
</tr>
<tr>
<td>Misspleling</td>
<td>False</td>
</tr>
</tbody>
</table>

Replace “misspleling” with the most sensible similar word.

“This sentence contains a misspelling.”
Determining word associations (e.g. finding novel drugs)

Basic steps:

1. Use regular expressions to match known words of interest

2. Define a context within which to consider each word
   - N-grams (similar to ‘neighborhood’ in real analysis)
   - Bag of words
   - Punctuation based (e.g. Which words were used in the same sentence)

3. Find other occurrences of contexts of interest:
   - Synonyms/antonyms – words that appear in similar contexts
   - Modifiers/adjectives – words that commonly appear around a word of interest are typically describing a characteristic of that word
Demonstration: An automated approach for classifying cause of fetal death

NLP FOR AUTOMATED ICD-10 CODING ASSIGNMENT
An automated approach for classifying cause of fetal death

Background
- NCHS provides cause of death coding for all death records in the United States including fetal deaths
- This predominantly manual approach takes time and resources to complete
- Upon receipt by NCHS, cause of death coding for fetal deaths can take years to complete

Objective: To create an automatic rule-based procedure for assignment of multiple cause ICD-10 codes for fetal death records at the national level
- Automating the classification of cause of death for fetal death records would provide an immediate benefit to research and surveillance efforts.
Data Source and Software

**Data Source:** 2014 – 2015 Fetal death reports

Literal text refers to the information written by the death certifier on the death report/certificate:

- Maternal Conditions/Diseases
- Complications of placenta, cord, or membranes
- Fetal Anomalies
- Injuries
- Infections
- Other field

Non-Literal information from the record includes:

- Weight of the fetus
- Plurality
- Length of gestation
- Sex

**Software:** R statistical language, focused on using base scripting language without additional resources
Project’s data flow

Spellchecking functionality
1. Dictionary
2. Distance metric
3. Language model

Abbreviation lookup table
- Abbreviation 1 -> meaning 1
- Abbreviation 2 -> meaning 2
- Abbreviation 3 -> meaning 3
  ...

Input text ➔ Spellchecking functionality ➔ Abbreviation lookup table ➔ Output text ➔ Classification Algorithm
COD Classification algorithm

*see D3 and Rshiny visualization* (or screenshots)
Defining topics for use by an algorithm

Syndromic and Case Definitions

- Fetal death related to Gestational Diabetes
- Flu symptoms at hospital intake
- Infant death from complication of methadone treatment
- Work related injuries
Defining a topic:

*Fetal death associated with gestational diabetes of the mother*
Defining a topic:

*Fetal death associated with complications of methadone therapy*
Representing all causes of death:
International Classification of Diseases
Text mining the coding manual to define all ICD codes
International Classification of Diseases (zoomed in)
International Classification of Diseases (high level view)

Shown here is just one subsection of one chapter of the ICD-10
shinyApp screenshots: *Fetal COD Classification algorithm*

**Getting Started**

- **What would you like to do?**
  - Choose from simple examples
  - Choose from complex examples
  - Explore the Spell Checking Functionality

- **Other (specify): ETPIC PREG.**

- **After some initial processing your text looks like this, and can be handed to the a.i.**
  - `mod control of gest dm dx at 17 wks`

**Generating ICD codes**

- **These are the ICD 10 codes that correspond to each section:**
  - **1. P01.0**
  - **1. P01.3**
  - **1. P70.0**
Getting Started

What would you like to do?
- Choose from simple examples
- Choose from complex examples
- Explore the Spell Checking Functionality

After some initial processing your text looks like this, and can be handed to the AI.

ectopic pregnancy polyhydramnios true knot in cord

gestational diabetes and placental abruption

press to run the algorithm

Understanding sentence structure

These examples are more complex sentences. The algorithm has broken your choice into the following sections:

1. ectopic pregnancy
2. polyhydramnios
3. true knot in cord

Generating ICD codes

These are the ICD 10 codes that correspond to each section:

1. P01.4
2. P01.3
3. P02.5

Understanding sentence structure

These examples are more complex sentences. The algorithm has broken your choice into the following sections:

1. gestational diabetes
2. placental abruption

Generating ICD codes

These are the ICD 10 codes that correspond to each section:

1. P01.1
2. P02.1

Understanding sentence structure

These examples are more complex sentences. The algorithm has broken your choice into the following sections:

1. premature preterm rupture of membranes
2. placental abruption

Generating ICD codes

These are the ICD 10 codes that correspond to each section:

1. P01.1
2. P02.1
Questions?

PATRICK DRAKE, STATISTICIAN
DIVISION OF VITAL STATISTICS | 301-458-4848 | NNH9@CDC.GOV