Feasibility and Utility of Lexical Analysis for Occupational Health Text

Philip Harber, MD, MPH and Gondy Leroy, PhD

Assessing Work–Asthma Interaction With Amazon Mechanical Turk
Philip Harber, MD, MPH and Gondy Leroy, PhD

Social media use for occupational lung disease

Public sharing of medical advice using social media: an analysis of Twitter
Gondy Leroy, PhD, Philip Harber, MD MPH, Debra Revere, MLIS, MA
Thank you for the opportunity to speak here.

“Hello there, my name is Phil Harber. I am currently speaking, but sometimes I see patients as a doctor in occupational medicine or in pulmonary medicine. I also play with computers and try to make them understand language sometimes. I am employed by a university and in a public health college but also go to the hospital sometimes. I sometimes do some administrative work and even traveled on aircraft yesterday. Last week, I coughed a little after I open my mouth doing a racing turn in the swimming pool.”

“I was welding on the stainless storage tank for the apple juice but my foot slipped and I had when in so I began to cough. The doctor said I might have asthma or maybe just a bad cold.”

We talk and think in words, not numbers
Using codes loses most information of prior slide:

- Good morning 19-1041
- 51-4120 111339 J45
- 786.2 541511 611310

- Language is rich, natural, flexible, and cheap.

- Codes are narrow, inter-relate poorly, and are not spoken by “real workers”.
Today’s Tools, Tomorrow’s Needs

- Storage: ~ Unlimited
- Computing Power:
- Data Sources: Many, mainly unstructured
- Programs- Rapidly evolving

...But....

- Interoperability
- Motivation to use

Limitations of Numerical Coding:

- Many jobs, industries, situations - poorly codable
- Only predesignated questions
- Highly structured input and highly structured output
- Very simple questions
- Modern Luddites
- Codes vs real life
Use #1: Autocoding (Structured $\rightarrow$ Structured)

Code: Replace a structured field with a structured number:

eg, Death certificate- Industry field $\rightarrow$ NAICS Code

**Auto coding**: Replace Nosologists with software

- Faster, cheaper, ? More accurate
- NIOSH NIOCCS, ...
#2: Abstraction

**Abstraction**

- Read the item
- Find the key term(s)
- Look up number
- Enter number in correct field
  - 😞 MSHA 7000 has 34 fields
- Needs domain knowledge
#3: Classifier

- Sorts items into one of a few alternative boxes
  - NIOSH trip/fall workers comp (Bertke)
  - CASCOT (UK)
  - DREAM (Harber)
  - SOCEYE (NCI)

- Numerous algorithms-identification, rule-based, probabilistic
- Most require **training set** and are **very specific**

4: Search

- Look for a specific term
- **Store raw words, not post-coded numbers**

- Does the public care about work-related asthma?
- Are ATS, AOEC, NIOSH outreach working ????

**TWITTER:**
- 40,000 tweets with “asthma”
- Very few involved a job, isocyanates,...
- Many, many, many included drugs, cam, children, air pollution, ozone,...

What we did...
5. Classify ALL items
6. Knowledge Discovery

- 86,000 MSHA incident reports
- Tokenize, Count words
- Eliminate irrelevant words (stop terms, low relevance)
- Annotate: assign to a domain
  - (e.g., exposure, injury/illness, medical care, action, ...)
- Count by domain
- Seek associations
What we did...
5. Classify **ALL** items
6. Knowledge **Discovery**

- **86,000** MSHA incident reports
- Tokenize, Count words
- Eliminate irrelevant words (stop terms, low relevance)
- Annotate: assign to a domain
  - (eg, exposure, injury/illness, medical care, action, ...)
- Count by domain
- Seek associations

Probabilistic
Bayesian classifiers
Syntactic analyzers
Pattern recognition
**Support Vector Machine**
Training set
Parser
Normalizer

→ **Lexical Analyzer**
Negation
Temporal relationships
Sentiment
Summary:

• Common Words (Injury/Illness Exposure/Hazard Treatment ...)

• Common causes

• Common effects

• Common cause-effect relationships
Associations: Cause & Effect

- **Automatic** discovery of RELATIONS between **HAZARDS** and **HEALTH EFFECTS**
**OUCH!!*#**

Targeted *a priori* example searches-

Start with either a hazard (Bolt) or an effect (Hand)

<table>
<thead>
<tr>
<th></th>
<th>BOLT</th>
<th>HAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>back</td>
<td>21%</td>
<td>rock</td>
</tr>
<tr>
<td>hand</td>
<td>9%</td>
<td>piece</td>
</tr>
<tr>
<td>arm</td>
<td>8%</td>
<td>roof</td>
</tr>
<tr>
<td>head</td>
<td>6%</td>
<td>belt</td>
</tr>
<tr>
<td>finger</td>
<td>5%</td>
<td>drill</td>
</tr>
<tr>
<td>rib</td>
<td>4%</td>
<td>steel</td>
</tr>
<tr>
<td>foot</td>
<td>4%</td>
<td>bolt</td>
</tr>
<tr>
<td>laceration</td>
<td>3%</td>
<td>miner</td>
</tr>
<tr>
<td>shoulder</td>
<td>3%</td>
<td>finger</td>
</tr>
<tr>
<td>pain</td>
<td>3%</td>
<td>metal</td>
</tr>
</tbody>
</table>
Annotation: Feasible & Specific
“How many words must I code to cover most info?”

- 1000 terms gets you > 80% coverage

- Topic-Specific Annotation
  - Exposure, technical term, health effect, medical care, activity, ...
  - Even easier if stems/ morphemes

- “KOMATSU”

Japanese construction equipment manufacturer, Komatsu is the name of the city in Japan where the company founded. Ko=Small, Matsu= Pine tree in Japanese. Komatsu has more than 250 patented inventions about sophisticated hydraulics.
DISCOVERY vs HYPOTHESIS DRIVEN

HYPOTHESIS DRIVEN

<table>
<thead>
<tr>
<th>BOLT</th>
<th>HAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>back</td>
<td>rock</td>
</tr>
<tr>
<td>hand</td>
<td>piece</td>
</tr>
<tr>
<td>arm</td>
<td>roof</td>
</tr>
<tr>
<td>head</td>
<td>belt</td>
</tr>
<tr>
<td>finger</td>
<td>drill</td>
</tr>
<tr>
<td>rib</td>
<td>steel</td>
</tr>
<tr>
<td>foot</td>
<td>bolt</td>
</tr>
<tr>
<td>laceration</td>
<td>miner</td>
</tr>
<tr>
<td>shoulder</td>
<td>finger</td>
</tr>
<tr>
<td>pain</td>
<td>metal</td>
</tr>
</tbody>
</table>

AUTOMATED KNOWLEDGE DISCOVERY

<table>
<thead>
<tr>
<th>CAUSE-EFFECT</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>drill</td>
<td>hand</td>
</tr>
<tr>
<td>belt</td>
<td>hand</td>
</tr>
<tr>
<td>hand</td>
<td>roof</td>
</tr>
<tr>
<td>bolt</td>
<td>hand</td>
</tr>
<tr>
<td>head</td>
<td>roof</td>
</tr>
<tr>
<td>belt</td>
<td>finger</td>
</tr>
</tbody>
</table>

P Harber  Univ of Arizona
# HYPOTHESIS DRIVEN vs DISCOVERY

## HYPOTHESIS DRIVEN

### Serendipity

<table>
<thead>
<tr>
<th>BOLT</th>
<th>HAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>back</td>
<td>rock</td>
</tr>
<tr>
<td>hand</td>
<td>piece</td>
</tr>
<tr>
<td>arm</td>
<td>roof</td>
</tr>
<tr>
<td>head</td>
<td>belt</td>
</tr>
<tr>
<td>finger</td>
<td>drill</td>
</tr>
<tr>
<td>rib</td>
<td>steel</td>
</tr>
<tr>
<td>foot</td>
<td>bolt</td>
</tr>
<tr>
<td>laceration</td>
<td>miner</td>
</tr>
<tr>
<td>shoulder</td>
<td>finger</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pain</th>
<th>3%</th>
</tr>
</thead>
</table>

## AUTOMATED KNOWLEDGE DISCOVERY

### Less biased

<table>
<thead>
<tr>
<th>CAUSE-EFFECT</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>drill</td>
<td>hand</td>
</tr>
<tr>
<td>belt</td>
<td>hand</td>
</tr>
<tr>
<td>hand</td>
<td>roof</td>
</tr>
<tr>
<td>bolt</td>
<td>hand</td>
</tr>
<tr>
<td>head</td>
<td>roof</td>
</tr>
<tr>
<td>belt</td>
<td>finger</td>
</tr>
</tbody>
</table>

- hammer
- roma
“Real language of work, workers...”

Gandy dancers, Santa Claus, Occupational physicians

Impersonates traditional holiday or storybook characters, such as Santa Claus, Snow White, and the Three Little Pigs, to promote sales activity in retail stores, at conventions or exhibits, and to amuse children at hospitals, amusement parks and private parties. Wears character costumes and impersonates characters portrayed to amuse children and adults. May hand out samples or presents, demonstrate toys, pose for pictures, and converse with

29-1071 Physician Assistants, Family Practice
29-1060 Physicians and Surgeons
29-1069 Physicians and Surgeons, All Other
29-1063 Physicians, Internal Medicine
29-1106 Physicians, Family Practice
29-1107 Physicians, Chiropractic
29-1108 Physicians, Naturopathic
29-1062 Physicians, Research
29-1064 Physicians, Surgery
Occupational health has its own language
(“Ontology”-Vocabulary, relations)

• Unified Medical Language System (UMLS) does not meet occupational health needs
• Medical terms → frequently in UMLS concepts, generally accurate
• Workplace terms → rarely present, usually wrong
WORDS- NOT NUMBER CODES 
FULL MEANING AND FLEXIBLE 
PEOPLE NOT PUNCH CARDS

THANK YOU!!

✓ Worker/patient talks to Automated Sir Val Ence (knighted epidemiologist)
✓ Who discovers new knowledge
✓ Provided by Marion (the automated librarian) to other workers