Seven Years of Machine Learning at the Bureau of Labor Statistics

Alexander Measure

Survey of Occupational Injuries and Illnesses

Example Narrative

Job title: sanitation worker

What was the employee doing just before the incident?
mopping floor in gym

What happened?
slipped on water on floor and fell

What part of the body was affected?
fractured right arm

What object directly harmed the employee?
wet floor

Codes Assigned

Occupation: 37-2011 (Janitor)

Nature: 111 (Fracture)

Part: 420 (Arm)

Event: 422 (Fall, slipping)

Source: 6620 (Floor)

Secondary: 9521(Water)

How should we attempt to automate it?

Rules

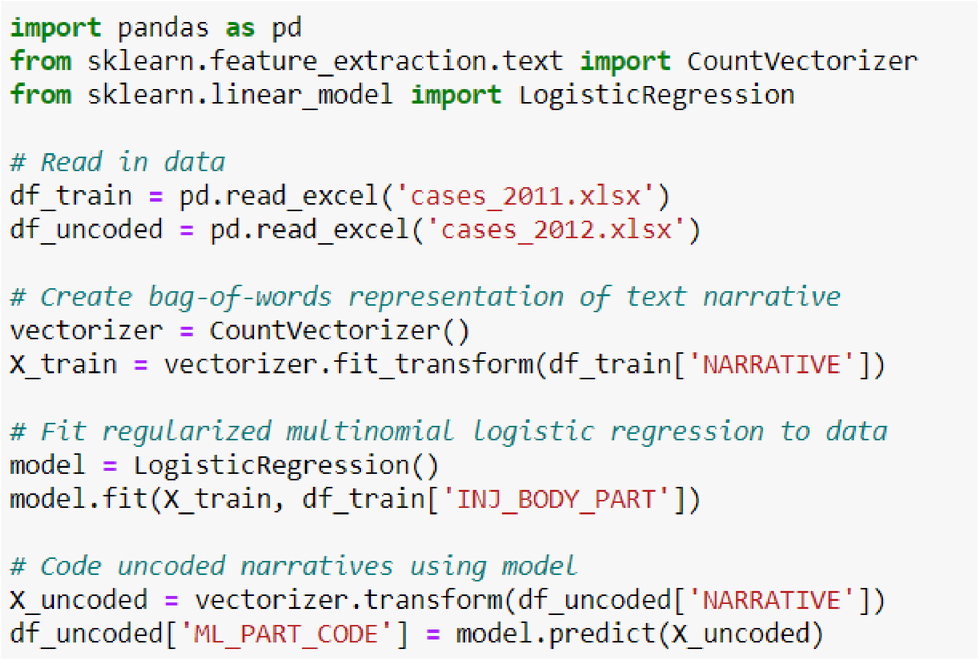
Job Title	Code
Janitor	37-2011
Environmental Services	37-2011
Senior janitor	37-2011
Cleaner	37-2011

Machine Learning

- Gather data
- Choose a model
- Fit to data

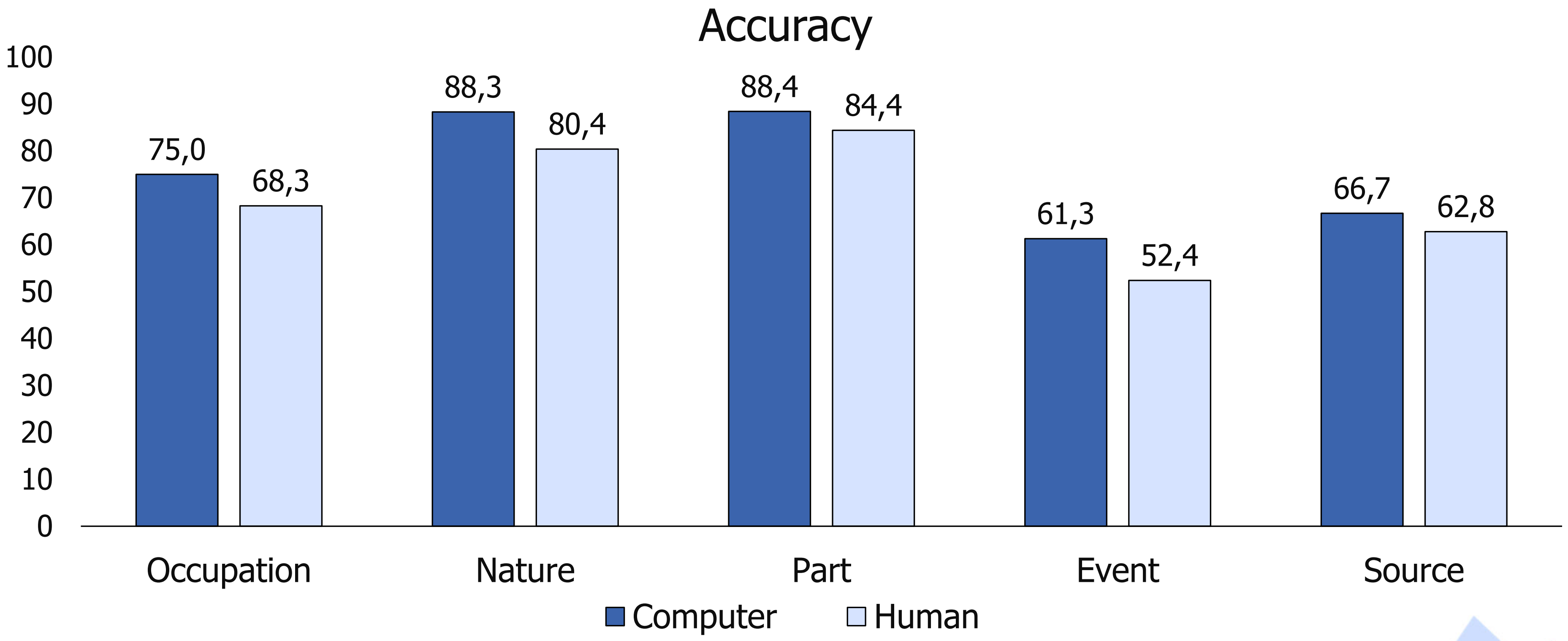
Simple enough to fit on a slide

- Bag-of-words
  - inputs are word occurrences
  - one for every word in training
- Regularized (L2)
  - reduces overfitting
- Multinomial
  - probability for every part code



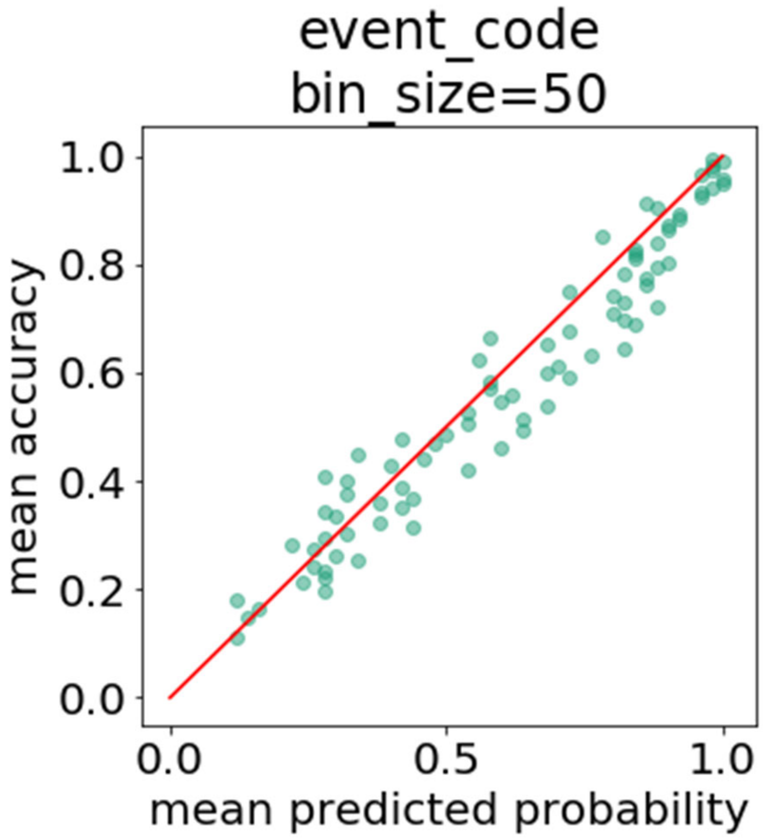
```python
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.linear_model import LogisticRegression

# Read in data
df_train = pd.read_excel('cases_2011.xlsx')
df_uncoded = pd.read_excel('cases_2012.xlsx')

# Create bag-of-words representation of text narrative
vectorizer = CountVectorizer()
X_train = vectorizer.fit_transform(df_train['NARRATIVE'])

# Fit regularized multinomial logistic regression to data
model = LogisticRegression()
model.fit(X_train, df_train['INJ_BODY_PART'])

# Code uncoded narratives using model
X_uncoded = vectorizer.transform(df_uncoded['NARRATIVE'])
df_uncoded['ML_PART_CODE'] = model.predict(X_uncoded)
```

Is it good enough?



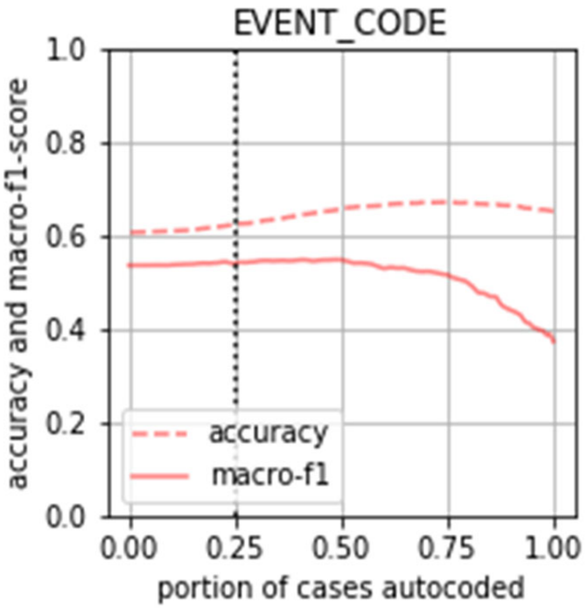
How should we use it?

- Accuracy isn't everything
- Predicted Prob ≈ True Prob
  - It mostly knows what it doesn't know
  - Humans can gather additional info
- But which threshold?

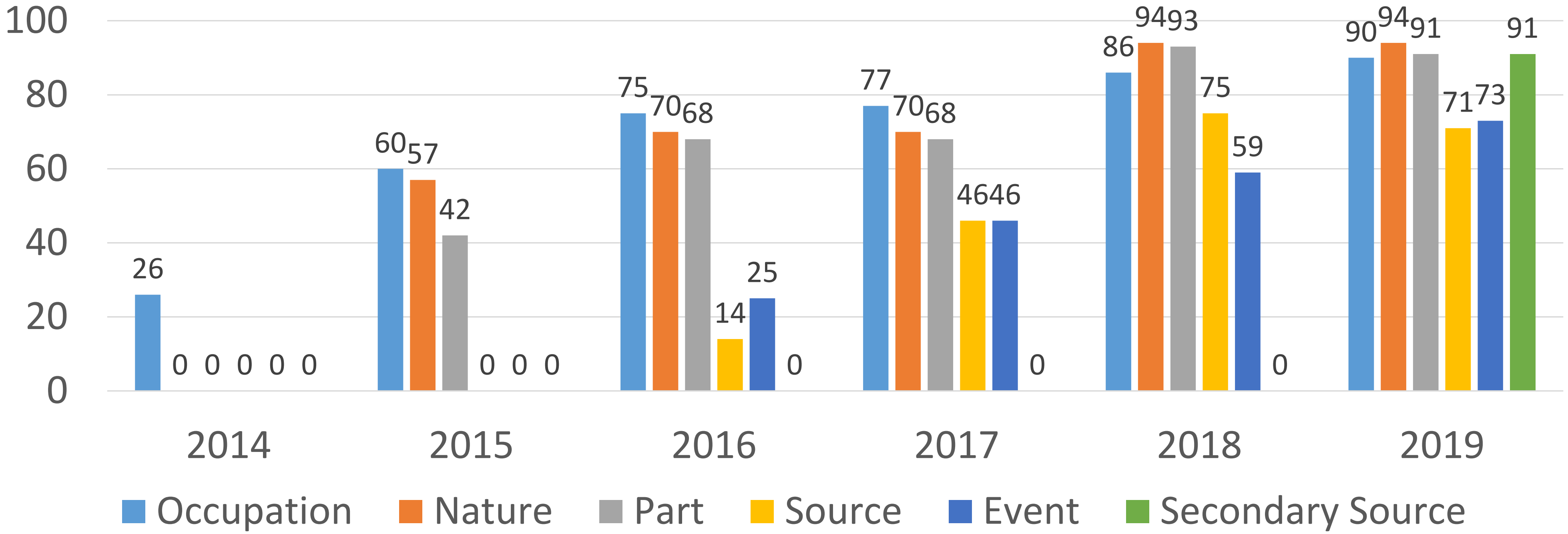


Finding the right balance

- Gold + Human + Computer codes allows simulation



Percentage of codes automatically assigned

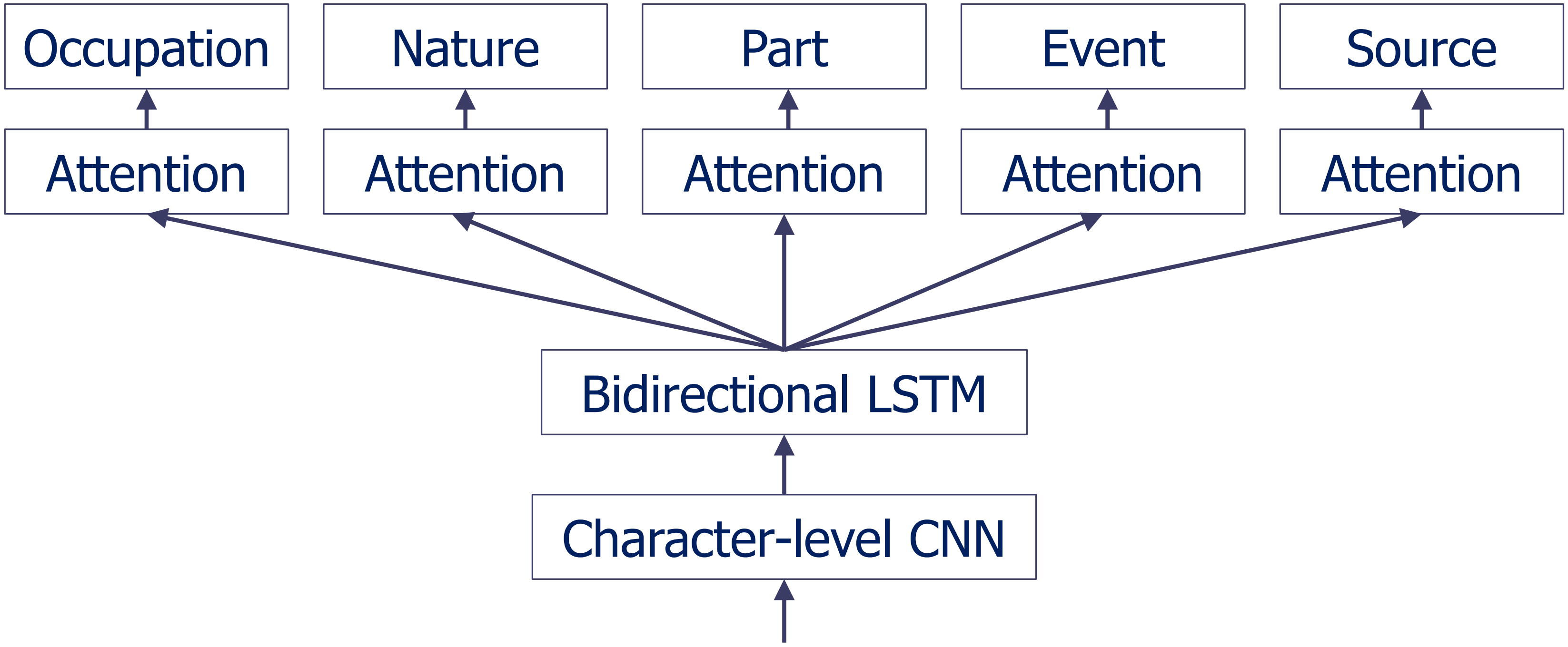


How do we maintain?

- How do we guard against unanticipated problems?
  - Before deployment: extensive evaluation
  - After deployment: human review

- How do we adapt to changes?
  - Gradual: retrain model each year
  - Sudden: human and rule-based intervention

How do we improve?



Occupation

Nature

Part

Event

Source

Attention

Attention

Attention

Attention

Attention

Bidirectional LSTM

Character-level CNN

<N0>janitor<N1>mopping floor in gym<N2>slipped on wet floor …

Contact Information

Alexander Measure

Occupational Safety and Health Statistics
www.bls.gov/iif/autocoding.htm
202-691-6185
measure.alex@bls.gov