An Introduction to Machine Learning and Natural Language Processing Tools

Presented by:
Nick Rizzolo, Mark Sammons, Vivek Srikumar, and James Clarke
What we need:

\[ f(\ ) = \text{“politician”} \]

\[ f(\ ) = \text{“athlete”} \]

\[ f(\ ) = \text{“CEO”} \]
Where to get it: Machine Learning

Data
- “politician”
- “athlete”
- “CEO”

Feature Functions

Learning Algorithm

→ f
Where Will We Get Data?

- What do we want it to look like?

- CNN corpus + NER tagger

Corporate Mogul
Athlete
Politician

...our way of life, " [LOC U.S.] President [PER Barack Obama] said in a written statement...
...assured [MISC Latino] immigration activists that President [PER Obama] will throw his support...
...neither [PER Obama] nor GOP presidential hopeful [PER John McCain] broached the topic...
How about features?

- Bag of word counts
- Filter the words by POS
  - Verbs may be important
  - Nouns, adjectives
- Other Named Entities
  - Who else is mentioned?

Politician

...our way of life, " [LOC U.S.] President [PER Barack Obama] said in a written statement...
...assured [MISC Latino] immigration activists that President [PER Obama] will throw his support...
...neither [PER Obama] nor GOP presidential hopeful [PER John McCain] broached the topic...

- Which words?
  - All words in the sentence
  - In a window around each mention
  - The closest with the POS we wanted
Implementation

- Design our own Entity class
  - Stores all sentences
  - Computes feature values

- Design a Parser
  - LBJ library interface
  - Returns labeled Entities one at a time

- Feature Engineering

- This is where the bulk of our time is spent