CS440
Natural Language Processing

Introduction to NLP

From Language to Information

• Automatically extract meaning and structure from:
  – Human language text and speech (news, social media, etc.)
  – Social networks
  – Genome sequences

• Interacting with humans via language
  – Smart speakers/dialog systems/chatbots
  – Question answering
NLP in industry

Information Retrieval

- 6,586,013,574 web searches every day (by one estimate)
- Text-based information retrieval is thus likely the most frequently used piece of software in the world
Text Classification: Disaster Response

- Haiti Earthquake 2010
- Classifying SMS messages

Mwen thomassin 32 nan pyron mwen ta renmen jwen yon ti dlo gras a dieu bo lakay mwen anfom se sel dlo nou bezwen

I am in Thomassin number 32, in the area named Pyron. I would like to have some water. Thank God we are fine, but we desperately need water.

Extracting Sentiment

- Lots of meaning is in connotation
  "connotation: an idea or feeling that a word invokes in addition to its literal or primary meaning."

- Extracting connotation is generally called sentiment analysis
Extracting social meaning from language

• Uncertainty (students in tutoring)
• Annoyance
  – callers to dialog systems
• Anger (police-community interaction)
• Deception
• Emotion
• Intoxication

Sentiment in Restaurant Reviews


900,000 Yelp reviews online

A very bad (one-star) review:

The bartender... absolutely horrible... we waited 10 min before we even got her attention... and then we had to wait 45 - FORTY FIVE! - minutes for our entrees… stalk the waitress to get the cheque… she didn't make eye contact or even break her stride to wait for a response…
What is the language of bad reviews?

- Negative sentiment language
  horrible, awful, terrible, bad, disgusting
- Past narratives about people
  waited, didn’t, was
  he, she, his, her,
  manager, customer, waitress, waiter
- Frequent mentions of we and us
  ... we were ignored until we flagged down a waiter to get our waitress ...

Computational Biology: Comparing Sequences

AGGCTATCACCTGACCTCCAGGCCGATGCCCTTAGCTATCACGACCggGTCGATTTGCCCCGAC

Sequence comparison is key to
- Finding genes
- Determining their function
- Uncovering evolutionary processes

This is also how spell checkers work!
Personal Assistants

Question Answering: IBM’s Watson
Why is language interpretation hard?

Ambiguity

- Resolving ambiguity is hard
Ambiguity

Find at least 5 meanings of this sentence:

I made her duck

• I cooked waterfowl for her benefit (to eat)
• I cooked waterfowl belonging to her
• I created the waterfowl statue she owns
• I caused her to quickly lower her head or body
• I recognized the true identity of her spy waterfowl
Ambiguity

I made her duck

Where is the ambiguity coming from?
- **Part of speech**: “duck” can be a noun or verb
- **Meaning**: “make” can mean “create” or “cook”

Ambiguity

**Grammar**: make can be:
- **Transitive**: (verb has a noun direct object)
  I cooked [waterfowl belonging to her]
- **Ditransitive**: (verb has 2 noun objects)
  I made [her] (into) [undifferentiated waterfowl]
- **Action-transitive** (verb has a direct object + verb)
  I caused [her] [to move her body]
Making progress on this problem…

• How we generally do this:
  – probabilistic models built from language data
    \[ P(“maison” \rightarrow “house”) \text{ high} \]
    \[ P(“L’avocat général” \rightarrow “the general avocado”) \text{ low} \]

Models and tools

• Language models
• Word embeddings
  – vector/neural models of meaning
• Machine Learning classifiers
  – Naïve Bayes
  – Logistic Regression
  – Neural Networks
Book

Speech and Language Processing (3rd ed. draft)
Dan Jurafsky and James H. Martin
https://web.stanford.edu/~jurafsky/slp3/

Data

Examples of interesting datasets...

The Dataset

- 6,685,900 reviews
- 192,609 businesses
- 200,000 pictures
- 10 metropolitan areas

1,223,094 tips by 1,037,138 users
Over 1.2 million business attributes like hours, parking, availability, and ambiance
Aggregated check-ins over time for each of the 192,609 businesses