COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Introduction to Natural Language Processing

CMSC 470

Marine Carpuat

Where we started on the 1st day of class

- Levels of linguistic analysis in NLP
  - Morphology, syntax, semantics, discourse
- Why is NLP hard?
  - Ambiguity
  - Sparse data
    - Zipf's law, corpus, word types and tokens
    - Variation and expressivity
  - Social Impact

Topics

- Words and their meanings
  - Distributional semantics and word sense disambiguation
  - Fundamentals of supervised classification
- Sequences
  - N-gram and neural language models
  - Sequence labeling tasks
  - Structured prediction and search algorithms
- Application: Machine Translation
- Trees
  - Syntax and grammars
  - Parsing

Ambiguity and Sparsity

- What are examples of NLP challenges due to ambiguity/sparsity?

- What are techniques for addressing ambiguity/sparsity in NLP systems?

Linguistic Knowledge

- How is linguistic knowledge incorporated in NLP systems?
    - Attention model as an example

NLP tasks often require predicting structured outputs

- What kind of output structures?
- Why is predicting structures challenging from a ML perspective?
- What techniques have we learned for addressing these challenges?

Structured prediction trade-offs in dependency parsing

Transition-based

- Locally trained
- Use greedy search algorithms
- Define features over a rich history of parsing decisions

Graph-based

- Globally trained
- Use exact (or near exact) search algorithms
- Define features over a limited history of parsing decisions

Structured prediction trade-offs in sequence labeling

Multiclass Classification at each time step

- Locally trained
- Make predictions greedily
- Can define features over history of tag predictions

Sequence labeling with structured perceptron

- Globally trained
- Use exact search algorithms
- Define features over a limited history of predictions

Consider this new NLP task

How would you build a system for this task?

- Goal: verify information using evidence from Wikipedia.
- Input: a factual claim involving one or more entities (resolvable to Wikipedia pages)
- Outputs:
  - the system must extract textual evidence (sets of sentences from Wikipedia pages) that support or refute the claim.
  - Using this evidence, label the claim as **Supported**, **Refuted** given the evidence or **NotEnoughInfo**.

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]

The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]

Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

This is the shared task of the Fact Extraction and Verification (FEVER) workshop

You can see what solutions researchers came up with here:

http://fever.ai/task.html

Social Impact

- NLP experiments and applications can have a direct effect on individual users' lives

- Some issues
  - Privacy
  - Exclusion
  - Overgeneralization
  - Dual-use problems

- What are examples of each of these issues in NLP systems?

[Hovy & Spruit ACL 2016]

Last few items

- Course
  - Project and final

- Keep learning
  - CLIP talks (Wed 11am) http://go.umd.edu/cliptalks
  - Language Science Center http://lsc.umd.edu
  - Podcasts:
    - NLP Highlights covers recent papers and trends in NLP research
    - Lingthusiam covers a very wide range of linguistic topics https://lingthusiasm.com/
    - Talking Machines: “Human Conversations about Machine Learning” https://www.thetalkingmachines.com