Introduction to NLP

CMSC 473/673 Spring 2017

Bryan Wilkinson

Course Overview

- Course Website is https://www.csee.umbc.edu/courses/undergraduate/473/
- Blackboard will be used for announcements and posting grades
- My Office Hours are in ITE 364
  - Tuesdays at 1PM
  - Wednesdays at 1PM
  - By appointment
- The TA is Aparna Subramanian
  - Thursdays at 2:30 PM
  - ITE 349

A Little About Me

- PhD Candidate in Computer Science
  - Planning to Defend in April
- I work with Dr. Tim Oates in CoRaL
- Research Projects done with the lab include:
  - Monitoring Twitter for Cybersecurity Attack Signals (w/ USNA)
  - Identifying Environmental Noise Sources in Recordings (w/ US Army Corps of Engineers)
- My dissertation research focuses on the semantics of adjectives
  - Ex: How can we learn what other adjectives modify the same property as *big* and *tiny*
  - To what degree do they modify that property?
- Other research interests include:
  - Working with endangered and under-resourced languages
  - Using artificial data

What is NLP?

- Natural Language Processing generally refers to the processing of text generated by humans for use in computation.
- Computational Linguistics is often used as a synonym but can also mean using computers to perform linguistic investigations or simulate linguistic theories.
- Distinction doesn’t really matter, tons of overlap between both

What is NLP?

- NLP **CAN** combine:
  - Computer Science
    - Information Retrieval
    - Machine Learning
    - AI
  - Math and Statistics
  - Linguistics
  - Philosophy
  - Literature
  - Pscyhology
  - Many more fields
- Many people practice NLP with only or mostly the first two.

Low Level Examples

- Changing a Verb's Tense
  - How do I make walk have past tense? What about catch?
- Parts of Speech
  - What is the noun in

The employee banks on getting a loan from the bank

- Grammar Agreement
  - Which is correct?

The students in my class (is | are) going to do great.

High Level Examples

- Automatic Speech Recognition
  - Assistants in Phones
  - Voice to Text
- Editing Assistance
  - What word was I trying to spell? THER
- Automatic Translation
  - ¿Por qué no puedo traducir a Wólof?
- Finding Abusive Text Online
  - How can we flag a posting for further review by a human?
- Where other applications can you think of?

So Why Is NLP Important

- A lot of things on previous slides might seem solved but….
- What about languages besides English?
  - Where do you even get the data?
- What happens if I am working with a new domain like medical text or tweets?
- How good are the current systems?
  - http://matrix.statmt.org/

Important Organizations and Conferences in NLP

- Association for Computational Linguistics (ACL)
  - Publishes *Computational Linguistics* journal
  - Holds ALC conference every year along with other local conferences (NAACL, EACL)
  - Has many special interest groups (SIGs) that focus on specific topics.
- International Committee on Computational Linguistics (ICCL)
  - Exists solely to plan COLING conference every two years
- European Language Resource Association
  - Organizes Language Resources and Evaluation Conference (LREC) every two years
  - Spearheaded creation of International Standard Language Resource Number (ISLRN)
- Linguistic Society of America (LSA)
  - Premier organization for all types of linguistics

What We Will Learn This Semester

- Wide breadth of different NLP areas
  - A little bit of Morphology, Syntax, Semantics, Pragmatics, and maybe some Phonology
- Applications using NLP
  - Translation, Summarization, Question Answering, etc.
- Some statistics
  - Needed for lots of NLP tasks and methods
- Some basic linguistics
  - Enough to understand what we are doing
  - And maybe inspire new ways of thinking about problems

What We Won't Learn This Semester

- Detailed Machine Learning Algorithms
  - Machine Learning has a lot of use in NLP
  - I'll give you the basics later today but we don't need any detailed knowledge
  - Don't need to implement standard Machine Learning Algorithms
- Neural Networks
  - Commonly used across all areas of Computer Science recently
  - I personally think it is better to understand problem thoroughly then apply tools rather than the other way around
  - That being said, I will try to point to relevant work using NNs when it is appropriate
- Complex Linguistic Theory
  - Not enough time to cover the intricacies of linguistics and teach NLP
  - UMBC offers minor in applied linguistics
  - I am happy to point out relevant courses if you are interested

Machine Learning Primer

- Where to Learn
  - CMSC 478/678
  - Lots of info all over the web
    - I like Course in Machine Learning (CIML) by Hal Daume III (NLP researcher at UMCP)
  - Don't need to implement standard Machine Learning Algorithms
- Algorithm Types
  - Supervised
    - Regression (Predict a numerical value from input)
    - Classification (Predict a class from input)
  - Unsupervised
    - Clustering
    - Dimensionality Reduction
- Awesome Libraries Exist
  - I am partial to scikit-learn (or sklearn) for Python

Classification

- The general idea behind classification is to assign a label to each “point” in the data
- Lots of good ways to do this
  - Support Vector Machines
  - Decision Trees
  - Neural Networks
  - Nearest Neighbors
- Useful in NLP
  - Given a sentence, label all the words with their [part of speech/ entity type/ semantic role]
  - Given a document, what is it about (Topic Modeling)
- For this class all you need to know is that classification produces some label as output when given a input
  - No need to know how it works (in general) or how to train it
  - We will look at sequence classification closely

Classification Example

- We want to know if *Zyrian* is a noun
  - Could do binary classification
  - Or multiclass classification
    - Equivalent to asking what part of speech is *Zyrian*
- What are some features that might help us decide?
  -
- Determine the value for these features for all words
- Then feed into some existing machine learning algorithm
  - Need a training set like words from the dictionary along with their part of speech to train the model

Dimensionality Reduction

- Instances in Machine Learning are often represented by large vectors of floating point numbers
  - Takes up a lot of space
  - Makes calculations take a while
  - Are all those dimensions needed?
- Dimensionality Reduction attempts to reduce the number of dimensions (features) needed to represent something
  - We will look at this when we talk about distributional semantics