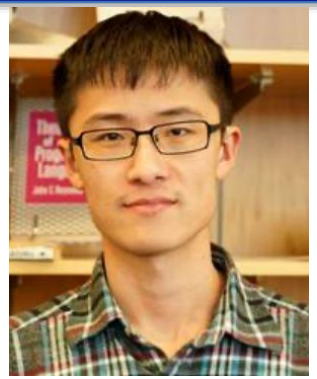
USC Viterbi School of Engineering
Department of Computer Science

CSCI 544
Applied Natural Language Processing

Mohammad Rostami
USC Computer Science Department

Who is Teaching This Course?

Mohammad Rostami



Xuezhe (Max) Ma

TAs

Sarik Ghazarian (head TA)

Jiao Sun

I-Hung Hsu

Ming-Chang Chiu

Graders

Vyshali Badanidyoor

Samipya Lahari Nirmal

Rucha Abhay Pande

Ujjwal Puri

Online Oriented

- Course Syllabus: check the link continuously

https://docs.google.com/document/d/1sMsAZ-Mid8c0hUlwzm7CVVp4hKMT3Q_BC0zqzWJ8-rE

- Course Blackboard: course material, quizzes & exams, announcement, assignments, etc.

https://blackboard.usc.edu/

- Course Zoom link:

https://usc.zoom.us/j/93834898825?pwd=dWRiaG1xS1g2c25YMEtqOUdVUVBzUT09

- Readings: available on syllabus and publicly available on the internet

- Read them before each lecture. Quizzes may include reading content that was not specifically covered in the class

- Course Slack Channel:

- fall21-csci-544-30249

Grading and Logistics

- Grading Schema

- Quizzes 10% (at the beginning of class on Tuesdays, from readings with focus on lectures. Only the top 10 quizzes will be considered.)
- Homework 40%
- Midterm: 20%
- Class Project 25% (Extra Credit)
- Final 5%

- Hard deadline: before start of the class
- Regarding requests: one week after announcing the grades (except final exam)

Grading and Logistics

- Collaboration

- Collaboration on HW is allowed but your code and report should be solely your own work. It is your responsibility to make sure your solution is not shared or copied by others.
- We will check for overlap in HW and treat it harshly
- You can use the office hours if things were not clear

- Pytorch Programing

- Expected to be self-studied
- A short introduction in the class

- Plagiarism

Course Timeline

Important dates
HW assignments: 8/26 to 11/9
Midterm: 10/26
Project Report: 12/02
Final Exam: 12/09

Course Project

- Groups of five:

- Start your groups early and email us
- Slack channel
- Randomly grouped

- Should have research and exploration nature (has not been done before but doable during the current semester):

- Your own topic, e.g., exploring your native language using existing tools, extending your prior work.
- Implementing and generating results for a good paper along with additional experiments to improve the paper
- Topics will be suggested
- Extra credit for interesting results

Applied Natural Language Processing

- Applied: we work with real language data
- Language: a structured system of communication based on signs, often in written form.
- What is a natural language?

- A means of communication that has evolved naturally in humans through use and repetition without conscious planning, e.g., English, Hindi, Chinese, Tamil, Arabic, etc.
- How about sign languages?

- What is **not** a natural language?

- Programing language: Python, etc.
- Formal languages: first order logic

Applied Natural Language Processing

- Processing: how to program computers to analyze large amounts of natural language data.

- **Natural language understanding**
- Natural language generation
- Speech understanding

- Applications

- Keyword Search
- Spell Checking
- Chatbot
- Machine Translation
- Dialogue Systems
- Grammarly!

Natural Language Levels of Representation

- Linguistics: the scientific study of natural language

Linguistics	NLP
Phonetics, Phonology	Speech recognition, synthesis
Morphology	Lemmatization, Stemming
Syntax	Part-of-speech tagging, Parsing
Lexical Semantics	Entity recognition
Compositional Semantics	Role labeling, Reference resolution Text classification
Production	Language generation Summarization Machine translation

Evolution of NLP

- Advancement of NLP

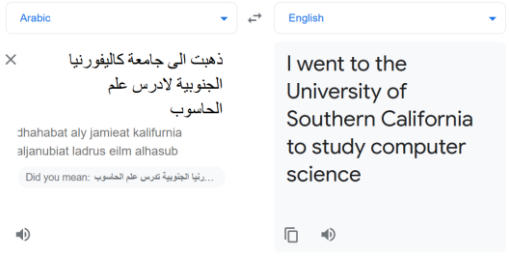
- Rule Based NLP:

  Who is the prime minister of India?

- NLP based on statistical learning: learning from extracted features

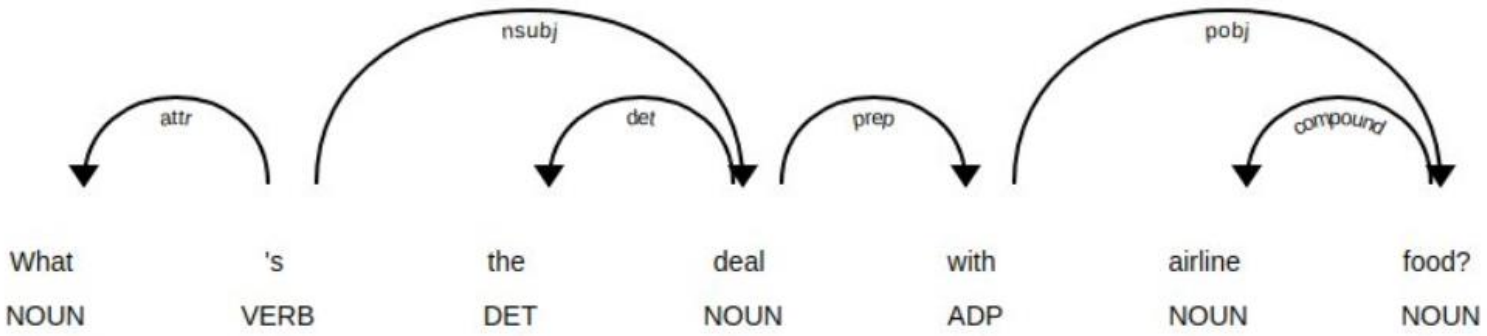
I don't like scary movies, but this movie is funny at the same time and that is why I liked it.

- Deep Learning: learning directly from large textual datasets



Rule-Based NLP

- A hand-crafted system of rules that parse text and match patterns is used to imitate the human way of building language structures.



- It can have high performance in specific use cases, e.g., question answering, but often suffer performance degradation when generalized
- Requires domain knowledge about the language

NLP Using Statistical Learning

- Learning is performed based on probabilistic modeling, likelihood maximization, and training classifiers.
- Requires an annotated training dataset along with a suitable method of feature engineering, potentially based on domain knowledge about the language.
- A parametric model is trained followed by evaluation on a test dataset similar, yet different, to the training dataset
- More generalizable than rule-based NLP and applicable to a broader range of applications, e.g., machine translation.

NLP Using Deep Learning

- Similar to machine learning with key differences:

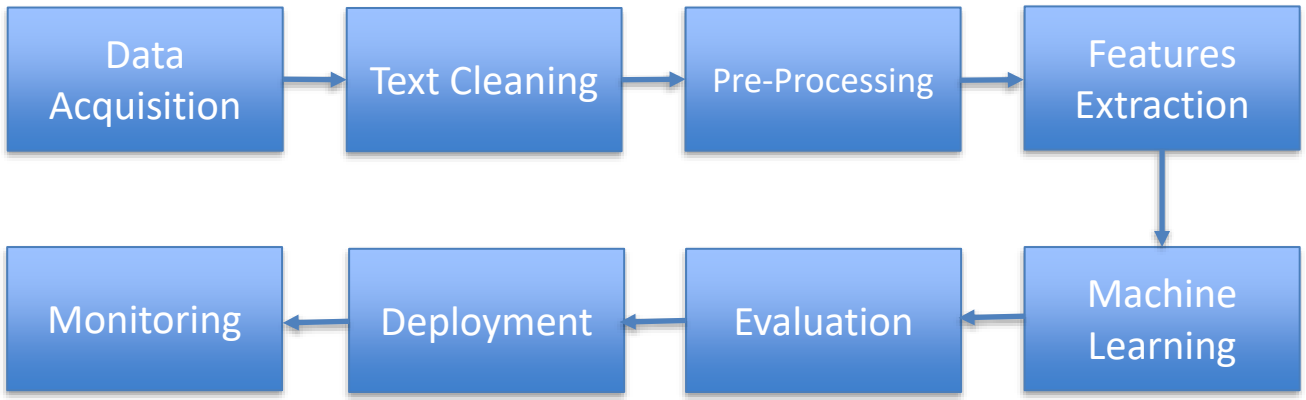
- Feature engineering is automated, as deep networks can learn important features from text -> domain knowledge about langue is minimized
- Raw text with minimal pre-processing is fed into the models
- It requires very large datasets to circumvent lack of using human domain knowledge
- Applicable to more challenging tasks, e.g., dialogue systems.

- Why do we still use rule-based NLP and traditional ML:

- Still good for sequence labeling
- Some ideas in deep learning are extended versions of earlier methods
- Can help to improve deep learning-based algorithms

NLP Pipeline

- We will mostly explore the stages between text cleaning and evaluation



Ambiguity in Natural Language

- Ambiguity
  - The shortstop caught the fly (lexical ambiguity)
  - Flying planes can be dangerous (structural/syntactic ambiguity)

- Local ambiguity vs global ambiguity: context
  - Fat people eat accumulates.
  - The man who hunts ducks out on weekends.

- Psycholinguistics: the study of human sentence processing

- Should NLP be done similar to human processing?

Preprocessing

- What is a word? Two words or one word?

- New York
- Hand-writing -> Handwriting
- He’s; It’ll -> He is; It will (constraction)
- David’s book, David’s happy
- Lower case

- **Tokenization**: splitting the text into units for processing

- Anywhere on the scale character to word
- Removing extra spaces
- Removing stop words
- Removing unhelpful words, e.g., external URL links
- Removing unhelpful characters, e.g., non-alphabetical characters.
- Ex: Tokenization is the first step in NLP. (7 tokens)

Preprocessing

- **Stemming:** wordform stripped of some characters
  - tokenization -> tokeniz
- **Lemmatization**: the base (or citation) form of a word
  - Tokenization, tokenize -> token
  - Went, gone, goes -> go
- Lemmatization can be lossy
  - EX: Where were you born? vs Where is your bear?
  - Is it important to be able to reconstruct the original?
- Implemented in the Python NLTK (assignment)
- Example:
  - I ordered these so I could fill up my Pandora bracelet right away, I didn't wanna have to wait for 10 more holidays before it was filled up. I liked the fact that they had a lot of pink in the charms, but barely got any in my shipment. I still have some of these charms on my bracelet but I don't wear it often since it turns my wrist green.
  - ordered could fill pandora bracelet right away wan na wait holiday filled liked fact lot pink charm barely got shipment still charm bracelet wear often since turn wrist green

Machine Learning

- Preprocessing + feature extraction converts input text into feature vectors

- Sentiment analysis:

I don’t like scary movies, but this movie is funny at the same time and that is why I liked it. -> Vector f with label +1 to denote positive sentiment

- Training a classifier using annotated features is a relatively solved problem in machine learning