Introduction to the Course

Prof. Sameer Singh

CS 295: STATISTICAL NLP
WINTER 2017

January 10, 2017

Based on slides from Nathan Schneider, Mohit Bansal, Sebastian Riedel, Yejin Choi, and everyone else they copied from.
About Me

Academic Positions

• New Assistant Professor at UC Irvine! (2016 -)
• Postdoc at University of Washington (2013 -)
• PhD from University of Massachusetts, Amherst (2014)

Research Interests

• **Natural Language Processing**: information extraction, relation extraction, entity linking and disambiguation, joint modeling
• **Machine Learning**: interpretable ML, semi-supervised learning, matrix/tensor factorization, probabilistic graphical models

http://sameersingh.org
sameer@uci.edu
Natural Language Processing

- Introduction to NLP
- Course Information
- Upcoming deadlines
Natural Language Processing

Introduction to NLP

Course Information

Upcoming deadlines
Knowledge Representation

Unstructured
Ambiguous
Lots and lots of it!
Humans can read them, but
... very slowly
... can’t remember all
... can’t answer questions

Structured
Precise, Actionable
Specific to the task
Computers can use
... quickly answer questions
... memory is not a problem
... don’t get tired
“Deep” understanding
Lots of Existing Applications

"You can ask for as many veggies as you want such as kimchi cucumber, crispy garlic and jalapeño, bean sprouts, etc. Then they give you sweet potato fries!" in 47 reviews

"I got the burrito with spicy pork and garlic shrimp, brown rice with gogi spicy sauce, broccoli, lettuce, kimchee cucumber and onions." in 64 reviews

"The stacked bowl had everything I love about Korean food, in a value priced $9ish dollars." in 20 reviews
But a long long way to go...

“You need to start understanding me Siri”

I’ll make a note of that.

“Yeah you better make a note of that”

Got it:

Of that
Future Applications

**Natural language processing in high demand**

Market expected to grow big time over next 5 years

By Bernie Monegain | August 14, 2015 | 10:14 AM

**Abstract**

**IMPORTANCE:** Natural language processing (NLP) has the potential to accelerate translation of cancer treatments from the laboratory to the clinic and will be a powerful tool in the era of personalized medicine. This technology can harvest important clinical variables trapped in the free-text narratives within electronic medical records.
Future Applications

Question Answering (instead of search)

Computational Social Sciences

Law, by reading past cases for you

Digital Humanities (historical texts)

Healthcare, by organizing records

Science, by reading papers for you

Assistive Technologies (dialog systems)

News Summarization
Turing’s test for Artificial Intelligence
Challenges in NLP

WHY ISN’T NLP SOLVED YET?
Three main challenges

- Ambiguity
- Sparsity
- Variation
Three main challenges

- Ambiguity
- Sparsity
- Variation
Language is Ambiguous

One tries to be as informative as one possibly can, and gives as much information as is needed, **and no more.**

- *Grice’s Maxim of Quantity*

**Corollary:** The more you know, the less you need.

Computers “know” very little.
Words have many meanings

Hershey’s Bars Protest
Words have many meanings

He knows you like your mother.
Attachment Ambiguities

Stolen painting found by tree.
Attachment Ambiguities

One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know.
- Groucho Marx
Attachment Ambiguities

She saw the man with the telescope.
And so on...

- Enraged Cow Injures Farmer with Ax
- Ban on Nude Dancing on Governor’s Desk
- Teacher Strikes Idle Kids
- Hospitals Are Sued by 7 Foot Doctors
- Iraqi Head Seeks Arms
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
Coreference Ambiguities

My girlfriend and I met my lawyer for a drink, but she became ill and had to leave.
Coreference Ambiguities

The city councilmen refused the demonstrators a permit because they feared violence.

The city councilmen refused the demonstrators a permit because they advocated violence.

“Context” is important

Winograd Schema: An Open Challenge for AI
Coreference Ambiguities
Entity Types and Identities

Types
• Washington, Georgia, Clinton, Adams
• John Deere, Williams, Dow Jones, Thomas Cook
• Princeton, Amazon, Kingston

Identities
• Same Name: Kevin Smith, Jamaica, Springfield
• Multiple “Names”: President, Obama, Chief, Bambam,...

“Context” is important
Animals with Misleading Names

Electric Eel
- Not an eel.

Mountain Goat
- Not a goat.

Maned Wolf
- Not a wolf.

King Cobra
- Not a cobra. Also, snakes are typically self-governing.

Peacock Mantis Shrimp
- Not a peacock.
- Not a mantis.
- Also, not a shrimp.

Horny Toad
- Not a toad.
- Only thinks of you as a friend.

Mayfly
- Active through the spring and summer.

Eastern Kingbird
- Found in the West.
- Many birds do not recognise its authority.

Not easy even for humans.
Three main challenges

- Ambiguity
- Sparsity
- Variation
Sparsity of Words

The graph shows the relationship between word frequency and rank. Words like "the", "of", "to", and "and" occur frequently, while words like "cornflakes", "mathematician", "fuzziness", "jumbling", "pseudo-rapporteur", "lobby-ridden", "perfunctorily", "Lycketoft", "UNCITRAL", "H-0695", "policyfor", and "Commissioneris" occur only once in the dataset.

>1/3
Sparsity of Words

Word frequency vs. rank

Frequency vs. Rank

0 20 40 60 80 100

0 200000 400000 600000 800000 1000000 1200000 1400000 1600000 1800000
Rescaling the Axes

Regardless of the size of the data, there will be many rare words.

Zipf’s Law

\[ f \times r = k \]

\[ \log f + \log r = \log k \]
In a document in which each character has been chosen randomly from a uniform distribution of all letters (plus a space character), the "words" follow the general trend of Zipf's. *(Try it at home!)*
Three main challenges

- Ambiguity
- Sparsity
- Variation
Many ways to say something

She gave the book to Tom vs. She gave Tom the book
Some kids popped by vs. A few children visited
Is that window still open? vs Please close the window
Variations in Domains

*Its vanished trees, the trees that had made way for Gatsby’s house, had once pandered in whispers to the last and greatest of all human dreams; for a transitory enchanted moment man must have held his breath in the presence of this continent, compelled into an aesthetic contemplation he neither understood nor desired, face to face for the last time in history with something commensurate to his capacity for wonder.*

ikr smh he asked fir yo last name so he can add u on fb lolololtw
Tools & Methods

HOW CAN WE GET COMPUTERS TO SOLVE THIS PROBLEM?
Sanders was born in Brooklyn, to Dorothy and Eli Sanders.
Two Different Approaches

DIRECTLY USE LINGUISTICS

Expensive, time-consuming...
... but also, incomplete!

MACHINE LEARNING!

Automatically learn from data!
... if the right data exists

“Every time I fire a linguist, my accuracy goes up.”
- Frederick Jelinek
Example: Machine Translation

Example: Machine Translation

Quiero ir a la playa más bonita.

Step 1: Break into Chunks
Example: Machine Translation

Quiero ir a la playa más bonita.

- I want  - to go  - to
- I love   - to work - at  - the beach
- I like   - to run  - per  - the seaside
- I try    - to appear - the open space
- I mean   - to be on - more pretty
-          - to - most pretty
-          - to - more lovely
-          - leave - most lovely
-          - pass away - more tidy
-          - forget - most tidy

Step 2: Translations for each chunk
Example: Machine Translation

Step 3: Generate all possible sequences

**Quiero ir a la playa más bonita.**

In same order
- I love | to leave | at | the seaside | more tidy.
- I mean | to be on | to | the open space | most lovely.
- I like | to be | on | per the seaside | more lovely.
- I mean | to go | to | the open space | most tidy.

In different order
- I try | to run | at | the prettiest | open space.
- I want | to run | per | the more tidy | open space.
- I mean | to forget | at | the tidiest | beach.
- I try | to go | per | the more tidy | seaside.

Step 4: Find the most human sounding one

I try | to leave | per | the most lovely | open space. 😞

I want | to go | to | the prettiest | beach. 😊

I want to go to the prettiest beach.
In summary...

Language to Knowledge
- Lots of applications...
- Made a lot of progress, but not done

It’s quite difficult
- Varied, sparse, and lots of ambiguities
- Context really matters

Machine Learning!
- With enough data and math, we can do it
- The future looks really exciting for NLP
Natural Language Processing

Introduction to NLP

Course Information

Upcoming deadlines
Course Logistics

Meetings

• Room: ICS 180
• Tues/Thursday 9:30-10:50
• No holidays this quarter (Yay!)

Reader

• Zhengli Zhao, PhD student
• Email: zhengliz@uci.edu
• But, contact us only on Piazza

Office Hours

• Room: DBH 4204
• Tuesdays 1pm - 5pm (by appt only)
• https://calendly.com/sameersingh/office-hours

Course webpage: http://sameersingh.org/courses/statnlp/wi17/
Learning Goals

Basics of NLP

• Familiarize you with NLP terms
• Tasks: Sequence Tagging, ...
• Methods: Neural approaches, ...
• Applications: Question Answering, ...
• Solve any NLP problem intelligently!

Critical Analysis

• Be able to read recent papers
• Appreciate their motivation
• Understand their approach
• Evaluate their results
• Can discuss ideas with NLP researchers!

Research Projects

• Be able to define a novel problem
• Study literature to identify overlap
• Implement existing and new methods
• Work in a team with researchers of different background
• With little guidance, have an NLP research agenda!
Topics (subject to change)

Words and Representations

- Text Classification: discriminative, generative, semi-supervised
- Word Vectors: vector semantics, dense embeddings, neural approaches

Language and Sequence Modeling

- Language Models: generative, discriminative, neural model
- Sequence Modeling: Part of speech and named entities, HMMs, CRFs

Sentence Structure Modeling

- Context free grammars, Probabilistic CFGs, constituent/dependency parsing
- Recursive neural models, sequence to sequence mapping, neural parsing

Applications and other topics

- Information Extraction: relations, coreference, entity linking, question answering
- Text generation, machine translation, entailment, reading comprehension, dialogs
Speech Recognition

HOW TO WRECK A NICE BEACH
Cognitive Sciences/Psycho-linguistics
Grading

Programming Homework 40%
Course Project 30%
Paper Summaries 15%
Participation 15%

Assignments
- All submissions through Canvas
- All deadlines are available now
- Will not be changing..
  - So start planning now

Late Submissions
- You get four *grace* days
  - Mention in the write-up
- Across all assignments
  - Use everyone’s for projects
- Full credit when used (no q asked)
- 0 if you run out (no partial credit)
Programming Assignments

4 Programming Assignments

- Throughout the quarter

Writing Up (PDF)

- Open-ended analysis of your approach
- Plots, figures, tables, examples...
- Think of it as a short research paper

Source Code (Python)

- Should be pretty straightforward
- Some skeleton Python code provided..
  - ..which you can ignore
- Piazza for potential bugs, weird results, etc.
Paper Summaries

3 Paper Summaries
• Due closer to the end of the quarter

Recent Conference Papers
• Cover all kinds of topics
• Randomly assigned to students
• You may not understand them!
  • But still have to summarize...

Summaries
• Content Summary: what they proposed
• Critical Analysis: what you liked/hated
• Instructions on the webpage already
Group Projects

Groups for the Project

• Ideal team size is 3, and diverse!
  • Absolute maximum of 4
  • <3 if I approve (ongoing work)

Submit Four Reports

• First two reports are very short (~1 page)
• Final report matters the most

Scope of Work

• Bigger the team, more ambitious the goal
• Has to be novel in some way
  • At least “workshop-level”
• Pitch and discuss ideas on Piazza

• More on projects in the next lecture..
Participation

Class participation
- Attend all the classes!
- Lectures should be discussions
  - Ask questions! Answer them!

Piazza participation
- Propose project ideas
- Ask/answer questions and issues
- Provide feedback to Instructor and TA
- Discuss readings and papers
Natural Language Processing

Introduction to NLP

Course Information

Upcoming deadlines
Upcoming...

Misc.
- Check out course webpage
- Check out Canvas, especially for deadlines
- Sign up for Piazza

Homework
- Homework 1 is up!
- Next two lectures will cover the topic
- Sign up for the Kaggle account (@uci.edu email)
- Due: January 26, 2017

Project
- Project pitch is due January 23, 2017!
- Start assembling teams now! (use Piazza)
- Start looking at papers, data, etc. for ideas