CSE 392 Natural Language Processing

Instructor: Yejin Choi

http://www.cs.stonybrook.edu/~ychoi/cse392
What is NLP?

- Artificial Intelligence dealing with human language.

- NLP is **AI-Complete**.

- **Turing Test**: Interrogator ‘c’ engages in a natural language conversation with ‘a’ and ‘b’ to determine which is a computer and which is a human.
What we say to dogs

Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!

What they hear

Blah blah GINGER blah blah blah blah GINGER blah blah blah blah...
Why is NLP hard?

Reason (1) – human language is ambiguous.

- Task: Pronoun Resolution
  - Jack drank the wine on the table. *It* was red and round.
  - Jack saw Sam at the party. *He* went back to the bar to get another drink.
  - Jack saw Sam at the party. *He* clearly had drunk too much.

[Adapted from Wilks (1975)]
Why is NLP hard?

Reason (1) – human language is ambiguous

- **Task**: Preposition Attachment (aka PP-attachment)

- I ate the bread **with** pecans.
- I ate the bread **with** fingers.
Why is NLP hard?

Reason (2) – requires reasoning beyond what is explicitly mentioned \((A,B)\), and some of the reasoning requires world knowledge \((C)\)

*I couldn’t submit my homework because my horse ate it.*

Implies that...

A. *I have a horse.*

B. *I did my homework.*

C. *My homework was done on a soft object (such as papers) as opposed to a hard/heavy object (such as a computer).*

– it’s more likely that my horse ate papers than a computer.
Why is NLP hard?

Reason (3) – Language is difficult even for human.

• Learning mother tongue (native language)
  -- you might think it’s easy, but…
    compare 5 year old V.S. 10 year old V.S. 20 year old

• Learning foreign languages
  – even harder
Is NLP really that hard?

In the back of your mind, if you’re still thinking...

“My native language is so easy. How hard can it be to type all the grammar rules, and idioms, etc into a software program? Sure it might take a while, but with enough people and money, it should be doable!”

You are not alone!
Brief History of NLP

- **Mid 1950’s – mid 1960’s: Birth of NLP and Linguistics**
  - At first, people thought NLP is easy! Researchers predicted that “machine translation” can be solved in 3 years or so.
  - Mostly hand-coded rules / linguistics-oriented approaches
  - The 3 year project continued for 10 years, but still no good result, despite the significant amount of expenditure.

- **Mid 1960’s – Mid 1970’s: A Dark Era**
  - After the initial hype, a dark era follows -- people started believing that machine translation is impossible, and most abandoned research for NLP.
Brief History of NLP

- 1970’s and early 1980’s – *Slow Revival of NLP*
  - Some research activities revived, but the emphasis is still on linguistically oriented, working on small toy problems with weak empirical evaluation

- Late 1980’s and 1990’s – *Statistical Revolution!*
  - By this time, the computing power increased substantially.
  - Data-driven, statistical approaches with simple representation win over complex hand-coded linguistic rules.

  ➔ “*Whenever I fire a linguist our machine translation performance improves.*” (Jelinek, 1988)

- 2000’s – *Statistics Powered by Linguistic Insights*
  - With more sophistication with the statistical models, richer linguistic representation starts finding a new value.
Why is NLP hard?

Reason (4) – representation v.s. computability

complex & rich representation

simple & plain representation

intractable

practical & tractable

linguistics

NLP
Why learn NLP?

• Because it’s fun.
  • It’s a field that is relatively young and growing rapidly
    => a lot of opportunities for being creative and making contributions.
Why learn NLP?

• Because you can make the world better.
  • Computer system that can help with your writing/composition
    • beyond spell checker or grammar checker

• Computer system that reads all the important blogs and news and provides you the summary
  • Product review analysis
Why learn NLP?

- Increase demands from industry
What to learn?

- Fundamental concepts and techniques
  - Language Models, Sequence Tagging, Trees

- Exposure to Research
  - How to read research papers
  - How to approach open-ended problems and make progress

- “A problem clearly stated is a problem half solved.”
  -- Dorthea Brande
Prerequisites

- Basic probability
- Basic statistics
- Basic linear algebra
- Machine learning helps *a lot*
- Algorithms
- Artificial Intelligence

- You must know how to code in some programming language
NLP Conferences to look out

- <ACL> Association for Computational Linguistics
- <NAACL> North American chapter of the ACL
- <EACL> European Chapter of the ACL
- <EMNLP> Empirical Methods in Natural Language Processing

- <CoNLL> Conference on Computational Natural Language Learning
- <COLING> International Conference on Computational Linguistics

- A Digital Archive of Research Papers in Computational Linguistics
Grading

- Exposure to research 15%
  - 3 sessions of research paper discussion
    (a) Written critiques (1 page)
    (b) Active participation in discussion sessions

- Homework 45%
  - 3 programming-oriented assignments

- Final Project 25%

- Quiz & Class Participation 15%.
Grading – Paper Discussion (15%)

1. Active Participation
   - One must attend all 3 sessions
   - Everyone must actively participate in each session. You can either make a comment, or ask a question, or answer a question.
   - It is recommended that you will bring an electronic or printed copy of the paper to the class.
Grading – Paper Discussion (15%)

2. Written Critique

- Submit the written critique at the beginning of each session as a printout, or email it to the instructor. Penalty if submitting your critique after the session.
- length: about 1 page
- suggested content:
  - summary of the paper in your own words
    - the summary portion should be less than 30% of your critique
    - word-to-word copy of a sentence to make a summary. Ok?
  - your own thought, criticism, suggestions, new research ideas
    - do not make empty statements: e.g., very interesting, I learned a lot
    - do not make trivial, shallow comments
Grading – Paper Discussion (15%)

2. Written Critique

- Submit the written critique at the beginning of each session as a print-out, or email it to the instructor. Penalty if submitting your critique after the session.
- length: about 1 page
- suggested content:
  - summary of the paper in your own words
    - the summary portion should be less than 30% of your critique
    - word-to-word copy of a sentence to make a summary. Ok?
      ➔ plagiarism! (Unless copied in quotes.)
      ➔ Re-word, rephrase everything in your own words
  - your own thought, criticism, suggestions, new research ideas
    - do not make empty statements: e.g., very interesting, I learned a lot
    - do not make trivial, shallow comments
Homework (45%)

- 3 programming-oriented assignments
- Some HW will be allowed to work in groups.
Exams

- No exam
- Occasional quizzes or in-class exercises.
Grading – Final Project (25%)

- Project proposal submission (5%)
  - Oct 18

- Project update presentation & submission (10%)
  - 10 minute presentation
  - Nov 11 (subject to change)

- Final project presentation & submission (15%)
  - 15 minute presentation
  - Dec 6 (fixed)
Grading – Final Project

• Work in groups (2~3 students)
• You may work alone only if
  • you can achieve at least 70% of what typical 2 people groups can achieve.
  • Most successful projects came out of students who worked in groups.
• Find partners no later than the end of Sep. If you need help in finding partners, send the instructor an email.
• Several project ideas will be given in the class. Or you can propose your own.
Successful past projects

**Learning General Connotation of Words using Graph-based Algorithms.**
Song Feng, Ritwik Bose, and Yejin Choi.  

**Domain Independent Authorship Attribution without Domain Adaptation.**
Rohith Menon and Yejin Choi.  
*Recent Advances in Natural Language Processing (RANLP)*, 2011.

**Composing Simple Image Descriptions using Web-scale N-grams.**
Siming Li, Girish Kulkarni, Tamara Berg, Alex Berg and Yejin Choi.  
*Computational Natural Language Learning (CoNLL)*, 2011.

**Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre.**
Ruchita Sarawgi, Kailash Gajulapalli and Yejin Choi.  
*Computational Natural Language Learning (CoNLL)*, 2011.

**Language of Vandalism: Improving Wikipedia Vandalism Detection via Stylometric Analysis.**
Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson and Yejin Choi.  
*Association for Computational Linguistics (ACL)*, short, 2011.
Late Submission

- Each student may adjust his/her homework deadline upto 7 days throughout the semester without a penalty. (not 7 days for each assignment, but 7 days cumulatively for the entire semester).
- Fractional values will be rounded up - for instance, late submission by 1 hour is counted as late by 1 day.
- After then, 10% of score will be subtracted each day.
- This rule does not apply to the critique submission.
- This policy is to encourage students to submit quality work, rather than poorly composed work in a hurry.
- If there are situations where the application of this rule can be ambiguous, I have the right to apply the rule as I see appropriate.
Textbook


- **[Library]** Some copies of the textbook are available in the reserve shelf of the North Reading Room (NRR) at the library.
- **[eBook]** Download NookStudy eBook platform and purchase the eBook license of the textbook at $62.10. Note that the license is valid only for one machine, and good for 180 days.
- **[Hard Copy]** The list price is $138.00, Amazon.com price is $104.27. Used books start at $88.99 as of Jan/22/2011. Apparently, you can sell your used book to Amazon.com at the end of the semester.
- **[Renting]** Stony Brook University Book Store (www.whywaitforbooks.com) offers Rental Program at $67.60.
- Go to Stony Brook University Book Store (www.whywaitforbooks.com) to explore above options
Reference Material

• Required text book:

• Other useful text book:
NLP 101: Syntax, Semantics, Pragmatics

- **Syntax** – grammatical ordering of words
- **Semantics** – meaning of words, phrases, sentences
- **Pragmatics** – meaning of words, phrases, sentences based on situational and social context
Syntax V.S. Semantics

(example by Noam Chomsky 1957)

- *Colorless green ideas sleep furiously.*
- *Furiously sleep ideas green colorless*
Semantics v.s. Pragmatics

What does "You have a green light" mean?

• You are holding a green light bulb?
• You have a green light to cross the street?
• You can go ahead with your plan?