CSE507 Computational Linguistics

Instructor: Yejin Choi

Class Homepage:

http://www.cs.stonybrook.edu/~ychoi/cse507
Course Description

- A general introduction to Computational Linguistics, the study of computational systems to understand and/or generate human language.
Computational **Linguistics** (CL)

- More historical name of the field. Often used with NLP interchangeably. Sometimes talks about more linguistically oriented part of the field.

National Language **Processing** (NLP)
National Language Processing (NLP) - CSE628
- language models, hmms, sequence tagging, syntax, information extraction, machine translation

Computational Linguistics (CL) – CSE507
- Semantics (lexical semantics, taxonomy, semantic parsing), pragmatics, discourse, coreference resolution, entailment
- +More focus on Natural Language Generation (summarization, sentence compression, dialog systems, computational poetry / story telling, image to text...)

➡️ overlap between 628 and 507:
about 3~4 lectures (out of 30 lectures).
Course Structure

- Fundamentals:
  - ~70% of the classes in the form of lectures
  - correspond to learning classic / fundamental topics in computational linguistics.

- Research components:
  - ~30% of the classes in the form of paper reading
  - correspond to learning recent advances in research & cutting-edge state-of-the-art.
    - About 5 sessions will be directly related to the fundamental components covered in the lectures.
    - About 5 sessions will explore topics in “natural language generation”.
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
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](http://www.whywaitforbooks.com)) offers Rental Program at $67.60.
- Go to Stony Brook University Book Store ([www.whywaitforbooks.com](http://www.whywaitforbooks.com)) to explore above options
Reference Material

- Required text book:

- Other useful text book:
Composition of Assignments

- 10 paper reading sessions:
  - Critique (upto 1 page)
  - Active participation
  - Each group of students (2 persons as a team) will be responsible for one presentation (upto 20 minutes) for one of the paper reading sessions.

- 2~3 programming homework

- final project
  - proposal report
  - update report + presentation (5 minute)
  - final report + presentation (10 minute)
Grading

- Paper discussion and critiques 20%
- Paper presentation 10%
- Homework 30%
- Final Project 30%
- Class Participation 10%
Paper Discussion and Critiques

- There will be 11 sessions --- you must attend at least 8 sessions.
- Everyone must actively participate in each session.
- Submit the written critique at the beginning of each session as a print-out. You will receive a penalty in your score if you walk in late for the paper discussion. You cannot submit your critique after the session. It is recommended that you will bring a printed copy of the paper to the class.

- Occasional Quizzes or Exams
Grading – Final Project

- Work in groups (2~3 students)
- You may work alone only if
  - you are a phd student
  - or 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 within the next two weeks. If you need help in finding partners, send me an email.
Grading – Final Project

- If I find that your final project is substantial enough to turn into a paper submission for reputable conferences, then your grade will become A automatically.
- If your project looks promising to turn into a publication, but not mature enough by the end of the semester, then you might get A- instead, depending on how you did on other part of the class.
- Above exceptions will not apply if
  - You worked on “potentially publishable ideas” and have some preliminary results, but the substance is still lacking, and the rest of your class performance is not good.
NLP Conferences to look out

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

- A Digital Archive of Research Papers in Computational Linguistics
Late submission:

- Each student may adjust his/her homework deadline up to 7 days throughout the semester without a penalty. (Not 7 days for each assignment, but 7 days cumulatively for the entire semester).
- This rule does not apply to “critique submission” for paper discussions.
- After then, 10% of score will be subtracted each day. This policy is to encourage students to submit quality work, rather than poorly composed work in a hurry.
- For the purpose of counting late submission, fractional values will be rounded up - for instance, late submission by 1 hour is counted as late by 1 day.
- 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. If you have a doubt, consult with me first before making your own assumption.
What is this class like

- Substantial project/programming assignments
- Not like traditional undergraduate classes
  - Heavy research paper reading (not easy)
  - No clean-cut problems and/or solutions
  - much more open-ended questions
  - Taste of original research
NLP is a young field

- Speech and Language Processing
  - First edition in 2000
  - Second edition in 2008
- Foundations in Statistical Natural Language Processing
  - First edition in 1999
Operating Systems

- Operating Systems Concepts
- 8th edition in 2011! (1st edition 3 decades ago!)
What is NLP/CL?

- 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 blah GINGER
blah blah blah
Why is NLP/CL hard?

Reason (1) – human language is ambiguous.

- Task: Pronoun Resolution (coreference 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/CL 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/CL 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/CL hard?

Reason (4) – representation v.s. computability

- Complex & rich representation
- Intractable
- Linguistics

- Simple & plain representation
- Practical & tractable
- NLP

- Computational Linguistics
Why learn NLP/CL?

- 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/CL?

- 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/CL?

- Because your future employer will love it.
Why learn NLP/CL?

- Think about what topic you would like to become an expert of for the next 20~40 years.
- Think about what field will prosper during the next 20~40 years.
  - Are there many interesting challenges yet to be solved?
  - Do you envision the field will make an impact on everyday human lives?
- Do you prefer a dynamic field where you are required to continually learn and have opportunities to innovate? (Or do you prefer a relatively static field.)
What to learn?

- Fundamental concepts and techniques
  - Focusing on Semantics, Pragmatics, and Natural Language Generation (NLG)

- Skills to embark on new research projects
  - How to read research papers
    - Useful even if you do not consider pursuing a ph.d. degree
  - How to approach unclear problems and make progress
    - Extremely useful both in industry and in research

- “A problem clearly stated is a problem half solved.”
  -- Dorthea Brande
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?
Please fill out answers for below questionnaire

1. Your name & email
2. Masters? Phd?
3. Planning to take the class? Or audit? Or haven’t decided?
4. Have you taken either “artificial intelligence” or “machine learning” or “NLP” class?
5. What is your area of interest outside NLP? e.g. systems, theory, etc