Course overview

LING 570

Fei Xia

Course web page

- Canvas page: https://canvas.uw.edu/courses/1264017

- Navigation menu:
  - Home: Office hours, links to zoom room, schedule, and course policy.
  - Announcement
  - Syllabus:
    - Prerequisites, textbooks, link to schedule
    - Course summary: generated automatically
  - Discussions
  - Assignments: assignment and submission
  - Grades
  - Conferences: for remote office hours
  - People: you can form study groups and get workspace for the group

Office hours

- Instructor: Fei Xia (fxia@uw.edu)
  - Office hours:
    - In-person: 2:30-3:30pm on Thurs (GUG 418A)
    - remote: 10-11am on Friday via Canvas

- TA: Sara Ng (sbng@uw.edu)
  - Office hours:
    - In-person: 2:15-3:15pm on Mon (GUG 416A)
    - remote: 1-2pm on Wed via Canvas

Course Prerequisites

- Programming Languages:
  - Java/C++/Python/Perl/..
- Operating Systems: Basic Unix/linux
- CS 326 (Data structures) or equivalent
  - Lists, trees, queues, stacks, hash tables, …
  - Sorting, searching, dynamic programming,..
- Automata, regular expressions,…
- Stat 390 or 391 (Probability and statistics): random variables, conditional probability, Bayes' rule, ….

LING 473 (summer 2018)

- Probability: random variables, independence, probability distributions
- Intro to Ling570 topics: FST, Formal grammar, LM, POS tagger, classifier, evaluation

➔ see http://courses.washington.edu/ling473/ (for summer 2017 version)

Textbooks

- Jurafsky and Martin, *Speech and Language Processing: An Introduction to NLP, Computational Linguistics, and Speech Recognition,* 2nd edition, 2009

- Manning and Schuetze, *Foundations of Statistical Natural Language Processing, 1999/2003.*
  - *UW library has eBook for the 1999 edition.*

- Several copies are available at the Treehouse lab. Please do not take them out of the lab.

Recording

- The recordings are available after class, posted in the “Tentative schedule”.

- Please remind me to
  - record the class
  - repeat the questions

- In-class students need to speak louder as the mic might not be very sensitive.

- The recordings and slides are all inside the schedule table.

Online Option

- The link to Zoom is on the home page:
  https://washington.zoom.us/my/clmsroom
- Online attendance for in-person students:
  - Not more than 3 times per course (e.g., bad weather)
- Please enter meeting room 5 mins before start of class
  - Try to stay online throughout class

Course workload

- Expect to spend 10-20 hours/week, including HW, reading, and class time.
- If you do not have so much for the course or cannot attend class live, you should wait and take the course later.
- Incomplete: only if all work is completed up last two weeks
  - UW policy

Course policy

- It is on the home page.
- Please read it carefully.
- The policy for ling571 and ling572 will be very similar (if not exactly the same).

Communication

- We will use Canvas:Announcement for important messages and reminders.
- When you email us, please use your UW email address or Canvas:
  - If you use your personal email address, please cc your UW email so that we know what your UW Netid is.
  - For questions that (potentially) require much discussion and/or clarification, email is not an effective way. Please ask in class or during office hours.
  - I check my UW emails more frequently than Canvas inbox. So for something urgent, use fxia@uw.edu.
- If you do not check Canvas often, please remember to set Account: Notifications in Canvas: e.g., “Notify me right away”, “send daily summary”.
- Do not send emails to the whole class except for emergency.
- For a non-urgent question, post to discussion board or ask in class / during office hours.

Programming languages

- Recommended languages:
  - C/C++/C#, Java, Python, Perl, Ruby, Mono, Jython
  - If you want to use a non-default version, use the correct path in your script.
  - See dropbox/18-19/570/languages for the file extension of those languages.

- If you want to choose a language that is NOT on that list:
  - You should contact Fei about this ASAP.
  - If the language is not currently supported on patas, it may take time to get that installed.
  - If your code does not run successfully, it could be hard for the grader to give partial credit for a language that (s)he is not familiar with.

Assignment due date

- All assignments are due at 11pm on Thurs (except that it is due on Wed for the week of Thanksgiving).

- No submission will be accepted 2 days after the due date.

Late penalty

- 1% for 1st hour, 10% for 1st 24 hours, 20% for 1st 48 hours.

- Ex: if the submission is 26 hours late, your score will be originalScore * 0.8.

Asking for extension

- You must contact Fei at least 24 hours in advance.
- Approved at the discretion of the instructor.
- Extension is at most 2 days.
- The submission area is closed after 2 days (regardless of whether you have an extension).
  - In other words, extension simply waives late penalty for the extension period.

Homework Submission

- For each assignment, submit two files through Canvas:
  - A note file: readme.txt or readme.pdf
  - A zipped tar file that includes everything: hw.tar.gz

    cd hwX/ # suppose hwX is your dir that includes all the files

    tar -czvf hw.tar.gz *

- Before submitting, run check_hwX.sh to check the tar file:

  /dropbox/18-19/570/hw1/check_hw1.sh hw.tar.gz

- check_hwX.sh checks only the existence of files, not the format or content of the files.

- For each shell script submitted, you also need to submit the source code and binary code: see 570/hwX/submit-file-list and 570/languages

Using patas

- You can debug code on your own machine, but you must test and run your final code on patas.
- The output files must be produced on patas.
- Your code will be tested on new data.

Regrading request

- You can request regrading for:
  - wrong submission or missing files: show the timestamp
  - crashed code that can be easily fixed (e.g., wrong version of compiler)
  - output files that are not produced on patas
- At most two requests for the course.
- 10% penalty for the part that is being regraded.
- For regrading and any other grade-related issues: you must contact TA within a week after the grade is made available.

Grading

- Grade:
  - Assignments: 100% (lowest score is removed)
  - Bonus for participation: up to 2%
  - The percentage is then mapped to final grade.
- No midterm or final exams
- Grades in Canvas:Grades
- TA feedback returned through Canvas:Assignments

Rubric

- Standard portion: 25 points
  - 2 points: hw.tar.gz submitted
  - 2 points: readme.[txt|pdf] submitted
  - 6 points: all files and folders are present in the expected location
  - 10 points: program runs to completion
  - 5 points: output of program on patas matches submitted output
- Assignment-specific portion: 75 points

Discussion board

- The 10-minute rule:
  - If you've been stuck on a problem for more than 10 minutes, post to the discussion board.

- The board is for student discussion only.
  - We (the instructor and TA) will not reply to individual posts directly.
  - For common issues, we will start a new thread and pin it.

- Tips:
  - Don't wait until the last minute to ask questions
  - Please read pinned threads. For others, you can decide whether you will read them.
  - For questions that are not resolved at the discussion board, please ask us during office hours or in class.

What is that question about?

Collaboration

- We encourage student collaboration: e.g., discussion board, study group, …
- Any kind of plagiarism is prohibited.
- If in doubt, please ask us first.

Course description

NLP applications

- Information retrieval
- Information extraction
- Question-answering
- Machine translation
- Dialog systems
- Sentiment Analysis
- …

Language & Intelligence

- Turing Test: (1949) – Operationalize intelligence
  - Two contestants: human, computer
  - Judge: human
  - Test: Interact via text questions
  - Question: Can you tell which contestant is human?

- Crucially requires language use and understanding

Knowledge of Language

- What does HAL (of 2001, A Space Odyssey) need to know to converse?

- *Dave: Open the pod bay doors, HAL.*
- *HAL: I'm sorry, Dave. I'm afraid I can't do that.*

- Phonetics/Phonology (LING550) and speech recognition:
  - Sounds of a language, acoustics
  - Legal sound sequences in words

Knowledge of Language

- What does HAL (of 2001, A Space Odyssey) need to know to converse?

- *Dave: Open the pod bay doors, HAL.*
- *HAL: I'm sorry, Dave. I'm afraid I can't do that.*

- Tokenization and Morphology (Ling 570)
  - Recognize, produce variation in word forms
  - Contraction: I'm → I am, can't → can not
  - Singular vs. plural: doors → /door/ + PL V
  - Verb inflection: am → /be/ + 1st person, sg, present

Knowledge of Language

- What does HAL (of 2001, A Space Odyssey) need to know to converse?

- *Dave: Open the pod bay doors, HAL.*
- *HAL: I'm sorry, Dave. I'm afraid I can't do that.*

- Part-of-speech tagging (Ling 570)
  - Identify word use in sentence
  - “can” is a verb, not a noun.

Knowledge of Language

- What does HAL (of 2001, A Space Odyssey) need to know to converse?

- *Dave: Open the pod bay doors, HAL.*
- *HAL: I'm sorry, Dave. I'm afraid I can't do that.*

- Syntax
  - Syntactic structure: subject, verb, object, etc.
  - Ling 566: analysis
  - Ling 570: chunking
  - Ling 571: parsing

Knowledge of Language

- What does HAL (of 2001, A Space Odyssey) need to know to converse?

- *Dave: Open the pod bay doors, HAL.*
- *HAL: I'm sorry, Dave. I'm afraid I can't do that.*

- Semantics (Ling 571)
  - The meaning of sentences:
    - individual (lexical), combined (compositional)
    - 'Open' : AGENT **cause** THEME to become *open*;

Knowledge of Language

- What does HAL (of 2001, A Space Odyssey) need to know to converse?

- *Dave: Open the pod bay doors, HAL*. (request)
- *HAL: I'm sorry, Dave. I'm afraid I can't do that.* (statement)

- Pragmatics/Discourse/Dialogue (Ling 571):
  - Interpret utterances in context
  - Speech act (request, statement)
  - Reference resolution: I = HAL; that = ‘open doors’
  - Politeness: I’m sorry, I’m afraid I can’t

Knowledge of Language

- What does HAL (of 2001, A Space Odyssey) need to know to converse?

- *Dave: Open the pod bay doors, HAL*. (request)
- *HAL: I'm sorry, Dave. I'm afraid I can't do that.* (statement)

- Others:
  - Various learning algorithms (ling572)
  - System design (ling573)
  - Speech recognition, dialogue, etc. (ling575)

Approaches to NLP

- Rule-based methods: before 1990s
- Statistical methods: 1990s to 2000s
- Deep learning: in the past decade

Shallow vs. Deep Processing

- Shallow processing (Ling 570):
  - Usually relies on surface forms (e.g., words)
    - Less elaborate linguistic representations
  - E.g. Part-of-speech tagging; Morphology; Chunking

- Deep processing (Ling 571):
  - Relies on more elaborate linguistic representations
    - Deep syntactic analysis (Parsing)
    - Rich spoken language understanding (NLU)

Topics covered in Ling570

- Unit #1: Introduction, probability theory (1 week)

- Unit #2: Formal language and FSA (1-2 weeks)
  - Formal language and formal grammar:
  - FSA
  - FST
  - Morphological analysis

- Unit #3: ngram models and HMM (2-3 weeks)
  - ngram LM and smoothing:
  - Part-of-speech (POS) tagging and HMM:
  - ngram tagger:

Topics covered in ling570 (cont)

- Unit #4: Classification (2-3 weeks)
  - Introduction to classification
  - POS tagging with classifiers
  - Chunking
  - Named-entity (NE) tagging

- Unit #5: NN intro, word embedding, neural LM (1-2 weeks)

- Unit #6: Other topics (1 week)
  - Information extraction (IE)
  - Summary

Coming up

- Get a patas account, if you don't have one.
- Go to canvas.uw.edu to check whether you have access to the course website.
  - If not, let me know by noon tomorrow.
- Next Tues:
  - Finish the quick review of probability theory.
  - Start on regular expression and FSA.
- Hw1 is available around 11pm today, and is due next Thurs at 11pm.